CHARACTERIZING DATA STREAMS OVER IEEE 802.11b

AD-HOC WIRELESS NETWORKS

THESIS

John T. Wagnon, 1Lt, USAF

AFIT/GIR/ENG/03-03

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GIR/ENG/03-03

CHARACTERIZING DATA STREAMS OVER IEEE 802.11b

AD-HOC WIRELESS NETWORKS

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science

John T. Wagnon, B.S.
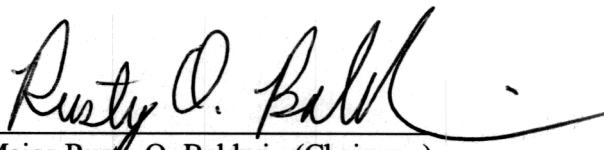
First Lieutenant, USAF

March 2003

AFIT/GIR/ENG/03-03

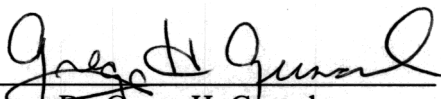CHARACTERIZING DATA STREAMS OVER IEEE 802.11b

AD-HOC WIRELESS NETWORKS

John T. Wagnon, B.S.
First Lieutenant, USAF

Approved:

Rusty O. Baldwin _____   _6 Mar 03_
Major Rusty O. Baldwin (Chairman)   date

Gregg H. Gunsch _____   _6 MAR 03_
Dr. Gregg H. Gunsch   date

Gilbert L. Peterson _____   _6 MAR 03_
Dr. Gilbert L. Peterson   date

Acknowledgements

Table of Contents

List of Figures

List of Tables

AFIT/GIR/ENG/03-03

Abstract

Soon, advancements in data encryption technology will make real-time decryption of the contents of network packets virtually impossible. This research anticipates this development and extracts useful information based on packet level characteristics. Distinguishing characteristics from e-mail, HTTP, print, and FTP applications are identified and analyzed. The analysis of collected data from an ad-hoc wireless network reveals that distinguishing characteristics of network traffic do indeed exist. These characteristics include packet size, packet frequency, inter-packet correlation, and channel utilization. Without knowing the contents of packets or the direction of the traffic flow, the applications accessing the wireless network can be determined.

CHARACTERIZING DATA STREAMS OVER IEEE 802.11b

AD-HOC WIRELESS NETWORKS

I. Introduction

Many studies have been conducted to detect abnormal activity on computer networks. These detection systems are created using techniques such as pattern recognition, data mining, statistical modeling, and Bayesian classification [SpZ00]. These techniques attempt to solve the problem of malicious activity in a reactive manner by capturing network traffic and determining in real time whether that traffic represents abnormal or malicious activity [GaH00]. This research effort is somewhat similar to these types of studies. This effort captures network traffic and determines what application is accessing the network without knowledge of the direction of the traffic flow or the contents of the packets.

1.1   Motivation

The one time pad encryption technique is considered an unbreakable code. This method of cryptography solves the problem of key distribution by generating enough random keys that the same key will never be used twice [Mul02]. This technique, along with others, supports the claim that data transmitted over computer networks will soon be unreadable [JaT99]. When the contents of a network packet are readable, the task of determining the type of application accessing the network is straightforward. The contents of the packet reveal what application is used. Due to the use of modern data encryption techniques, other methods of characterizing data exchange among users on a

computer network are needed. Characteristics of network traffic not affected by encryption must be explored.

1.2   Characteristics of Network Traffic

The IEEE 802.11 standard for wireless networks is extremely popular for wireless use [CWK96]. Packets in an IEEE 802.11b ad-hoc wireless network have characteristics that are likely to distinguish one application from another. These packet characteristics include 1) packet size, 2) packet frequency, and 3) inter-packet correlation. Packet size, measured in bytes, is simply the total length of the transmitted packet. Packet frequency refers to the number of times a certain packet size is transmitted during a given time interval. Inter-packet correlation considers patterns that arise from certain transmissions. Signal power and channel utilization are also studied as distinguishing characteristics of network traffic.

The following scenario illustrates how these characteristics can characterize a certain data exchange in an ad-hoc wireless network. Suppose a user sends an e-mail and the size of every second packet is 1500 bytes. Other packets are 500 bytes in size. A user accesses a web page and the size of every fifth packet is 1200 bytes while the remaining packets are 500 bytes. Or, a user prints a document at a remote printer and the size of the first two out of three packets is 800 bytes while the third packet is 1500 bytes. These patterns repeat until all data transmits for the various applications. Each of these notional examples highlights a dominant characteristic that can be recognized. Packet size assists in distinguishing among these three applications because the only application that creates packet sizes of 800 bytes is the printer application, the only application that creates packet sizes of 1200 bytes is the HTTP application, and the only application that

creates packet sizes of 1500 bytes is the e-mail application. Packet frequency is

important because half the e-mail traffic consists of packets sized 1500 bytes while the

other half are packets sized 500 bytes. The HTTP traffic has the same 500 byte packets

as the e-mail traffic, but the frequency of 500 byte packets in the HTTP traffic is much

greater because four out of every five packets in this HTTP traffic is sized 500 bytes.

Simply observing the frequency of packets of size 500 bytes could lead to the

characterization of the data exchange. Inter-packet correlation can be seen in the printer

traffic since the pattern of packet sizes is medium, medium, long where a short packet is

less than 600 bytes, a medium packet is 600 to 900 bytes and a long packet is larger than

900 bytes. Using this same technique, the e-mail traffic would produce a packet pattern

of short, long, short, long and the HTTP traffic would produce a packet pattern of short,

short, short, short, long. These three characteristics of network packets do not depend on

knowing the contents of the packet, yet they can be used to characterize the data

exchange on a network.

Other characteristics of the network such as signal power and channel utilization

are also studied. Signal power is used to estimate the distance of a wireless device from

the observation point. If a certain application is used only when two machines are within

a certain radius of one another, signal power "threshold" could be postulated, thus

eliminating the possibility of that application being used on the network. This knowledge

could also eliminate false positive determinations of applications being used. Channel

utilization could also be important because applications could utilize a given channel at

different rates. If this is true, utilization measurements may be sufficient to determine

what application is accessing the network.

When characteristics like packet size, packet frequency, inter-packet correlation, signal power, and channel utilization are measured and analyzed, several outcomes are possible. These include success, false positives, false negatives, and unknown. Success occurs when the correct application is identified by the chosen method of data characterization. False positives occur when an application is determined to be used on the network when in fact that application is not used. False negatives are cases where applications are determined as not being used on the network when in fact they are. Unknowns occur when a sample of network traffic is captured and analyzed and the application cannot be determined. The false positives, false negatives, and unknowns are to be avoided while the success outcome is acceptable.

1.3 Summary

This chapter discusses the background and motivation for this research. A discussion of how encryption techniques have influenced network monitoring is provided. Additionally, a new approach of analyzing packet characteristics is discussed. Chapter II reviews relevant literature supporting the research. Chapter III outlines the methodology used to conduct the research. Chapter IV provides the analysis of the collected data. Finally, Chapter V presents concluding remarks, limitations of the research, and recommendations for future study in this area.

II. Literature Review

2.1 Background

Recent advancements in the technologies of data encryption have made monitoring network packets at the bit level very difficult; soon it may be nearly impossible [Mul02]. Examining network packets at the bit level is computationally intensive even when the information contained in the packet is not encrypted. When the bits of a data packet are encrypted, the contents are unreadable [JaT99]. Monitoring, then, becomes infeasible and other methods of detecting the nature of the network traffic are required. This chapter includes an in-depth literature review of techniques used to monitor network packets.

The following section describes previous research in the area of data characterization. Various techniques to study the content of network traffic are then discussed. Although bits within a packet can be encrypted, the packets themselves are organized and built according to various protocols. Because of this organization, opportunities exist for determining network traffic content at the packet level.

2.2 Previous Related Research

Several studies characterize data in various environments. These studies provide support and motivation for this study. A network to recognize hand gestures has been developed [HuH97]. Using a Fourier descriptor, one hand figure is distinguished from another. By matching input gesture models to previously stored models, different hand gestures are characterized. Using a genetic research technique, foreign patterns in computer network traffic have been recognized [DaG02]. In a genetic search, novel pattern detectors evolve and varying degrees of abnormality of network traffic are

distinguished. Using these techniques on network traffic, abnormal network traffic is characterized and identified.

Classification techniques have been used to characterize contours found in images [CWP02]. A set of characteristics is extracted from the image and used to distinguish sharp contours from blurred contours in the image. A method to quantify generic features of angles of the lower body extremities of human subjects has also been proposed [Lak00]. Salient features are extracted from the angles of the hip, knee, and ankle to characterize normal and pathological walking. Research that is motivated by the "cocktail party" effect is a recent development. This effect describes the phenomenon in which we are able to focus on one voice in an environment where many voices are heard simultaneously. A system that uses characteristics of a single voice to successfully identify a specific speaker in the presence of many competing speakers has been developed [PMS00]. Data mining and pattern recognition techniques to distinguish circular tornadic weather patterns from other types of weather patterns have been used by [TWF00]. The data characteristics of radar reflectivity, mean Doppler velocity, and spectrum width are used to compare known tornadic activity to unknown weather activity. Using these characteristics, unknown weather patterns are classified as tornadic or not. An algorithm to determine the 3D orientation of an aircraft uses a set of spatial moments as a feature to characterize the view of the aircraft [AgC98]. An unknown set of moments is compared to known 3D orientation moments, and the orientation of the aircraft is thereby determined.

The research above has goals similar to this research. Characteristics are extracted and used to distinguish one thing from another. Although many of the

techniques may not apply to computer networks, they do support the idea of distinguishing one application from another used in this research. The sections below describe several techniques used to study wireless network traffic and distinguish one application from another.

2.3   Data Mining

Data mining is a process which automatically or semi-automatically discovers patterns in data. Various methods are used to detect structural patterns in data so that explanations and predictions can be made. Several different techniques are used to accomplish data mining. The following sections address the techniques of statistical modeling and naïve Bayesian classifiers, decision trees, and covering algorithms.

2.3.1   Statistical Modeling and Naïve Bayesian Classifiers

When evaluating data, a simple technique known as statistical modeling can be used to make decisions about patterns. The following example is a slightly modified version of an example found in [WiF00]. Table 2-1 includes fictitious network data but is used to illustrate the principles of many different data mining techniques. The instances in this dataset are characterized by two different attributes, packet size and packet frequency. The outcome is a "normal network traffic" or "abnormal network traffic" decision.

Packet size is measured in bytes, packet frequency indicates how many times a packet of size n, where n is a fixed number of bytes, occurs within a given time interval t. The time interval can be set to any value but for the purposes of this example, t is equal to one second.

Table 2-1 Sample Network Data

Packet Size (in bytes)	Packet Frequency (t = 1 sec)	Normal
Small	Low	No
Small	High	Yes
Large	High	No
Medium	Low	No
Medium	Medium	Yes
Small	Medium	No
Large	Low	Yes

Whether the network traffic is normal or not depends on the values of each of the

attributes in the table. The table lists packet size as being small, medium, or large. Small

packet size is defined as a packet of less than 200 bytes. Medium packet size is defined

as a packet of less than 1,001 bytes but more than 200 bytes. Large packet size is defined

as a packet of more than 1,000 bytes. Packet frequency is listed in the table as being low,

medium or high. Low packet frequency is defined as a packet of size n that occurs less

than 11 times during a one second interval. Medium frequency is defined as a packet of

size n that occurs less than 21 times but more than 10 times during a one second interval.

High packet frequency is defined as a packet of size n that occurs more than 20 times

during a one second time interval. The values determining the outcome of each of the

different attributes are completely notional and do not actually represent true network

traffic. Regardless of the accuracy of the attribute values, the concept of statistical

modeling and Bayesian classification can be shown.

From the data represented in Table 2-1, certain statistical models can be built.

Table 2-2 takes the data recorded in Table 2-1 and shows the number of times each

attribute-value pair occurs and the associated outcome for that pair. For example,

Table 2-2 records that when a packet size is small, Normal occurs once. The Normal column is slightly different than the other two columns because it simply counts the occurrences of yes and no.

Table 2-2 Network Data With Counts

Packet Size (in bytes)			Packet Frequency (t = 1 sec)			Normal	
Normal?	Yes	No	Normal?	Yes	No	Yes	No
Small	1	2	Low	1	2		
Medium	1	1	Medium	1	1	3	4
Large	1	1	High	1	1		

Table 2-3 records the probabilities associated with each attribute-value pair. For example, the probability of the data being Normal given the packet size is small is 1/3.

Table 2-3 Network Data With Probabilities

Packet Size (in bytes)			Packet Frequency (t = 1 sec)			Normal	
Normal?	Yes	No	Normal?	Yes	No	Yes	No
Small	1/3	2/4	Low	1/3	2/4		
Medium	1/3	1/4	Medium	1/3	1/4	3/7	4/7
Large	1/3	1/4	High	1/3	1/4		

The packet size and packet frequency data shown in the tables is assumed to be independent. This is one area of concern noted by critics of the statistical modeling method of data mining [WiF00]. Although few things in this world are independent of

something else, this approach in statistical modeling still yields surprisingly accurate results [TiZ99]. Given the data shown in Tables 2-2 and 2-3, certain predictions are made concerning future network traffic. Consider the following example in Table 2-4.

Table 2-4 Data Collected From A Different Time Interval

Packet Size (in bytes)	**Packet Frequency (t = 1)**	**Normal**
Large	Medium	?

Because the original table (Table 2-1) does not have a complete representation of all possible combinations of attributes and outcomes and because the new data shown in Table 2-4 reveals a combination of attributes not yet seen, statistical modeling can be used to determine if the traffic is normal or not. Treating the two features of Tables 2-2 and 2-3 (packet size and packet frequency) as independent events allows the probabilities of those events to be multiplied together. This can be used to calculate the likelihood of *Normal = yes* and *Normal = no* using the attributes in Table 2-4 and the probabilities in Table 2-3:

$$\text{Likelihood of yes} = 1/3 * 1/3 * 3/7 = \mathbf{0.0476} \qquad (2\text{-}1)$$

$$\text{Likelihood of no} = 1/4 * 1/4 * 4/7 = \mathbf{0.0357} \qquad (2\text{-}2)$$

Because Table 2-4 shows that packet size is large and packet frequency is medium, the likelihood of yes value is found by multiplying the probabilities found in Table 2-3. The probability of data being normal given the packet size is large is 1/3. The probability of data being normal given the packet frequency is medium is 1/3. The probability of data being normal is 3/7. This process is repeated for the value of likelihood of no. The

2-6

results of these equations show that the outcome of the new data is more likely to be

normal than abnormal. By normalizing the two numbers, they can be turned into

probabilities

$$\text{Probability of yes} = \frac{.0476}{.0476 + .0357} = 57.14\% \qquad (2\text{-}3)$$

$$\text{Probability of no} = \frac{.0357}{.0476 + .0357} = 42.86\% \ . \qquad (2\text{-}4)$$

This method of calculating probabilities of events is based on Bayes' rule of conditional

probability [WiF00]. Bayes' rule states that, given a hypothesis H, and evidence E which

bears on the hypothesis, then $\Pr[H \mid E] = \dfrac{\Pr[E \mid H] * \Pr[H]}{\Pr[E]}$. In the previous example,

the hypothesis H is the data is normal given the evidence that packet size is large and

packet frequency is medium. The attributes of packet size and packet frequency can be

labeled as evidence E_1 and E_2 respectively. Applying Bayes' formula, the following

results are obtained

$$\Pr[normal = yes \mid E_1, E_2] = \frac{\Pr[E_1 \mid yes] * \Pr[E_2 \mid yes] * \Pr[yes]}{\Pr[E_1] * \Pr[E_2]} \ . \qquad (2\text{-}5)$$

The probability of the evidence, $\Pr[E_1] * \Pr[E_2]$, can be ignored because it will go away

when the results are normalized. When the values are inserted for the conditional

probabilities in the equation above, the result is the same as the previous result,

yes = 57.14% and no = 42.86%. This method is referred to as Naïve Bayesian

Classification because it assumes "naively" that the evidence attributes are independent.

Although this independence may not be true in many cases, the formula still works

surprisingly well [WiF00]. Using the results from the equations above, the answer to

Table 2-4 is *Normal = yes* because yes has a probability of 57.14% while no has a

probability of 42.86%. This same result is shown using the decision tree approach in the next section.

2.3.2   *Decision Trees*

Another technique used in data mining is decision trees. The recursive construction of decision trees typically includes selecting an attribute to place at the root node with a branch for every possible value of that attribute. Each leaf value can then be similarly evaluated and more branches can be built from those new nodes. Determining which attribute to split on is a critical decision [KaP00]. Figure 2-1 shows decision tree stumps built from the data in Table 2-1. This type of structure is the building block for the decision tree analysis.
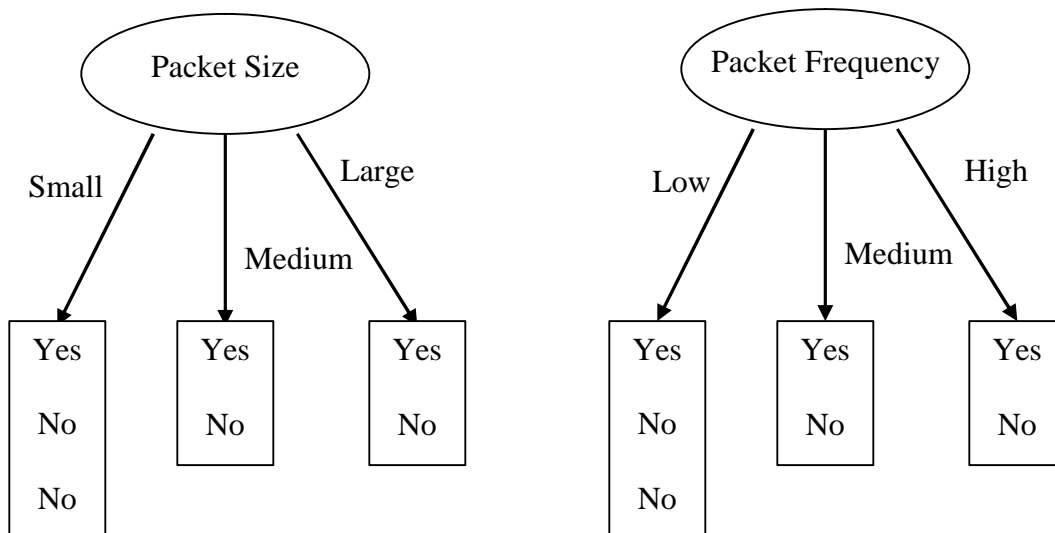
Figure 2-1 Decision Tree Stumps

After the decision tree stumps have been built, a branch is chosen based on values called *entropy*. These values range from zero to one and represent the amount of information needed to decide whether a given branch is the best choice. In the example above, the

entropy would help determine whether a new data set is normal network data or not. The representation of entropy follows certain syntax. For the packet size node above, the entropy for each branch (small, medium and large) is represented as entropy([1,2]), entropy([1,1]) and entropy([1,1]) respectively. When the number of either yes's or no's is zero, the entropy is zero. When the number of yes's and no's is equal, the entropy is one. The value equation for entropy must also hold for multiclass situations where more than two values are present. To calculate the entropy of a set and adhere to the rules concerning entropy, the following equation is used

$$\text{entropy}([p_1, p_2, p_3, \ldots, p_n]) = -p_1 * \log p_1 - p_2 * \log p_2 - p_3 * \log p_3 - \ldots - p_n * \log p_n \quad (2\text{-}6)$$

where p_1, p_2,…,p_n are the fractions of values in the leaves of the tree and always sum to one. The logarithms used in equation 2-6 are base 2. Since the sum of the values is always one, the entropy will always be between zero and one. Because the logarithm of a fraction is a negative number, the leading coefficient must also be negative so that the resulting entropy remains positive. The logarithm of zero is mathematically undefined, but for the purposes of this calculation, it is assumed to be a very small number. The result of the multiplication is then defined and equal to zero. Evaluating the entropy for the packet size node results in

$$\text{entropy}([1,2]) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.918 \qquad (2\text{-}7)$$

$$\text{entropy}([1,1]) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1.0 \qquad (2\text{-}8)$$

$$\text{entropy}([1,1]) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1.0 \,. \qquad (2\text{-}9)$$

Notice that when the same number of yes's and no's are present in the leaf of a branch, the entropy is equal to one. Once the entropy for each leaf has been determined, the entropy for the node can be calculated. This is done by multiplying the entropy of each leaf by the ratio of instances of each leaf to the total number of instances for the node. After multiplying each entropy value with its associated fraction, the results are all summed to give the total nodal entropy. The total nodal entropy for the packet size node is

$$\text{entropy}([1,2], [1,1], [1,1]) = 0.918 \times \frac{3}{7} + 1 \times \frac{2}{7} + 1 \times \frac{2}{7} = 0.965. \qquad (2\text{-}10)$$

Before any of the branches were created for the packet size node, the number of yes's and no's were three and four respectively. In order to find out how much information was gained by evaluating the branches of a node, the entropy is computed for the entire data set. The data set entropy is

$$\text{entropy}([3,4]) = -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right) = 0.985. \qquad (2\text{-}11)$$

By subtracting the average entropy of each branch of the packet size node by the entropy of the data set, the information gain for that node can be determined. The information gain for the packet size node is

$$\text{gain(packet size)} = \text{entropy}([3,4]) - \text{entropy}([1,2], [1,1], [1,1]) \qquad (2\text{-}12)$$

$$= 0.985 - 0.965$$

$$= 0.02$$

Information gain values can be similarly calculated for each of the nodes. All information gain values are listed in Table 2-5.

Table 2-5 Gain Values for Each Node in the Decision Tree

gain(packet size)	0.02
gain(packet frequency)	0.02

Normally, one node will have a greater gain value than the others, and the node with the greatest gain value is chosen. When the gain values are equal, it does not matter which one is chosen next. Typically a top down left to right method is used in the case of equal gain values, so the packet size will be used to complete this example [RuN95]. The node chosen after the first iteration of calculations becomes the splitting node for the first split. The remaining attributes are then evaluated as a branch of the three values for packet size. This is illustrated in Figure 2-2.
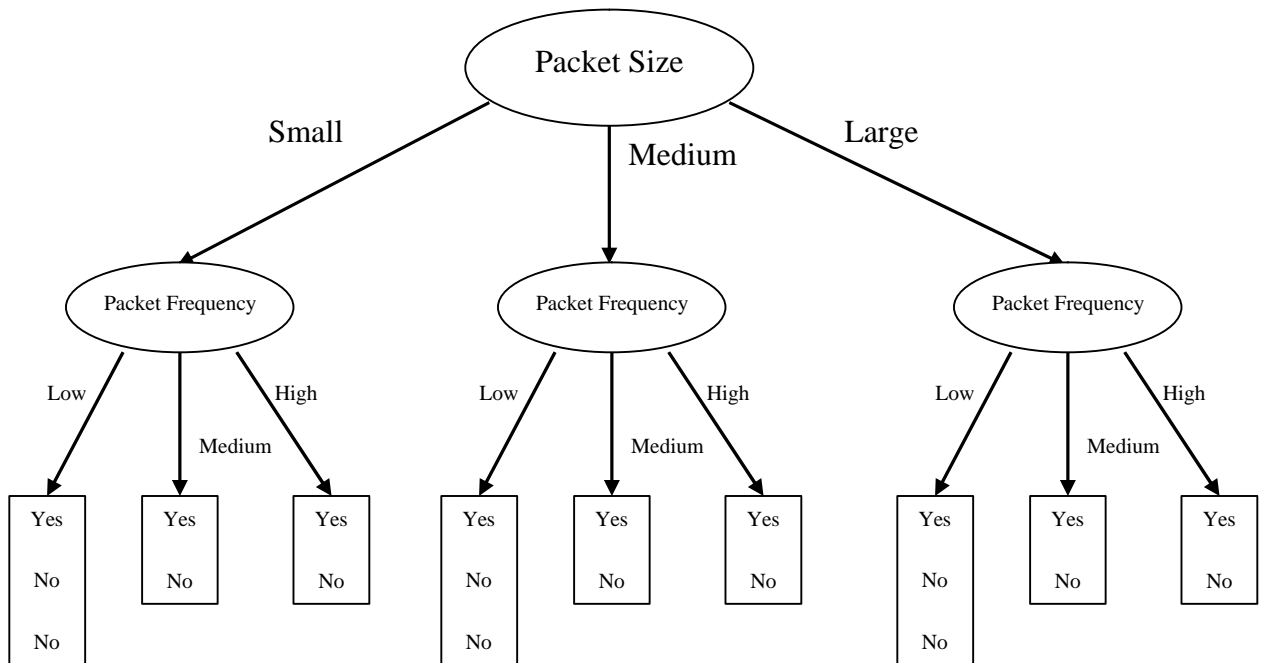
Figure 2-2 Expanded Tree Stumps For Network Data

After the tree is expanded, the entropy for each new leaf can be calculated. The entropy

is recalculated for each leaf until the entropy for two consecutive calculations is the same.

When this happens for all leaves, the tree is finished. Ideally, each leaf will have an

entropy of zero, but that is not always the case. When a new data set is presented, it

simply follows the path of the decision tree and the value for the network traffic will

either be normal or not normal. A final decision tree is shown in Figure 2-3. Notice that

the leaves of each branch simply have a yes or a no and represent the value for *Normal*

given that path.
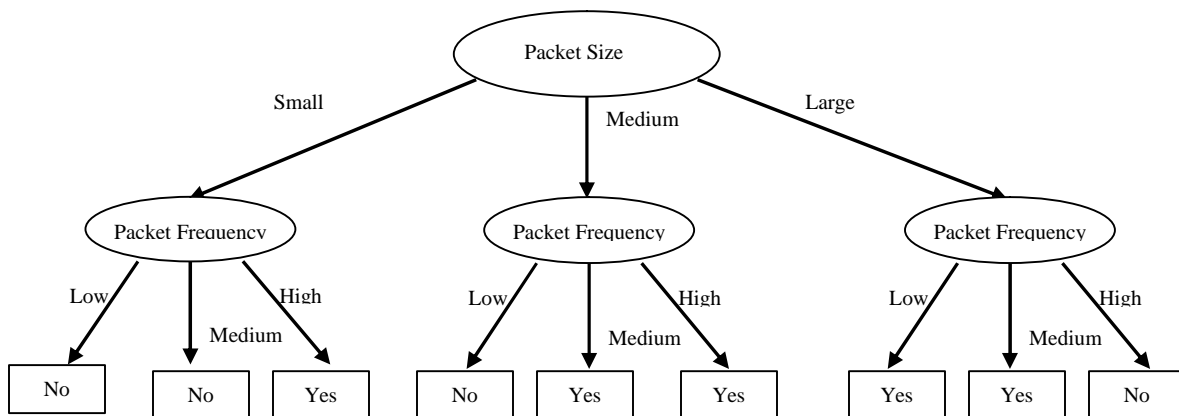
Figure 2-3 Final Decision Tree For Network Data

2.3.3 *Covering Algorithms*

While decision trees use a divide and conquer approach to solving problems,

covering algorithms attempt to "cover" certain values of a problem set while excluding

values not in the set. Instead of a decision tree as the result, the covering algorithm leads

to a set of rules that explain the data. The following example is due to [WiF00] and

shows a simple example of what a covering algorithm accomplishes. The graph in Figure

2-4 shows the total set of data being considered.

In this example, the *a*'s will be covered first. This leads to the formation of the

first rule: if x > 1.5 then class = *a* (cf., Figure 2-5(a)). Clearly, this will eliminate many

of the instances of *b* but not all of them.
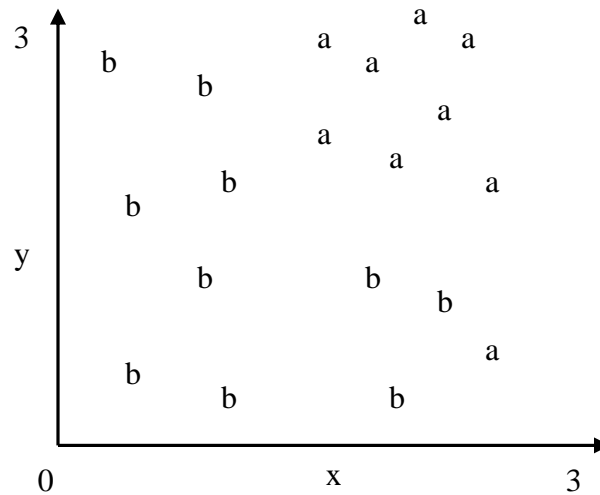
Figure 2-4 Data Set For a Covering Algorithm

For this reason, another rule needs to be included: if x > 1.5 and y > 1.6 then class = *a*

(cf., Figure 2-5(b)). When these rules are implemented, all of the *a*'s but one are

covered. If it becomes necessary to include the remaining *a*, additional rules can be

implemented. Obviously this would increase the complexity of the problem, so a

decision must be made balancing the covering of the data and the simplicity of the rules.

When the rules are determined for covering the *a*'s, this implies rules for the coverage of

the *b*'s. If x ≤ 1.5 then class = *b* and, if x > 1.5 and y ≤ 1.6 then class = *b*. Figure 2-5 (a)

shows the graph after the first rule has been applied and Figure 2-5 (b) shows the graph

after the second rule has been applied.

Covering algorithms work similar to decision trees. They apply one rule to try to cover as many instances as possible and then apply additional rules until an acceptable number of instances are covered [BHM01]. Decision trees split the nodes of the tree and rely on information gain values in order to maximize the separation of classes.
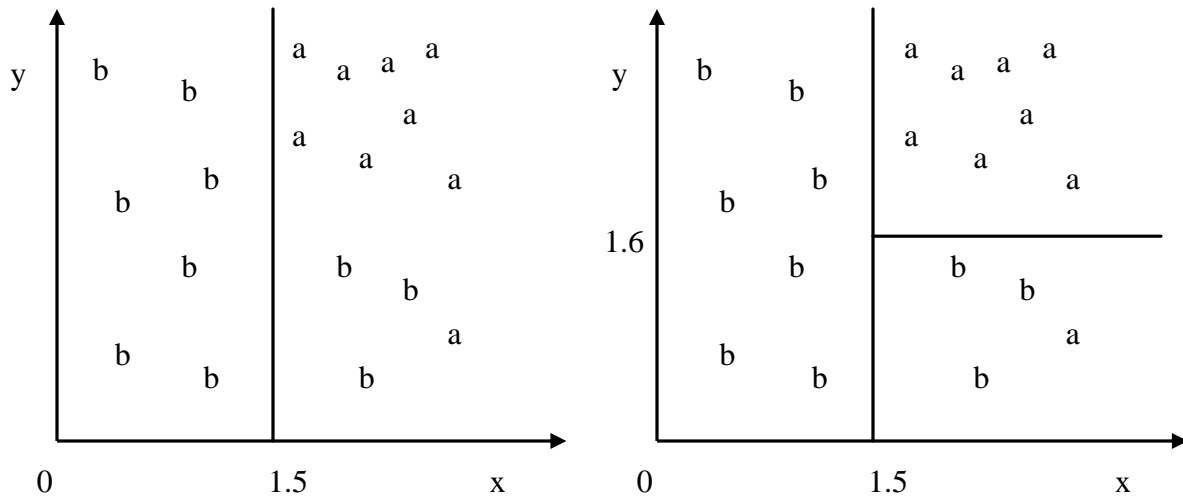
Figure 2-5 (a) Graph of Data After Rule 1   **Figure 2-5 (b)** Graph of Data After Rule 2

Covering algorithms attempt to add rules to data in order to maximize accuracy of coverage without regard to the separation of classes. They attempt to choose an attribute-value pair that maximizes the probability of a desired classification [WiF00].

2.4   Pattern Recognition

Pattern recognition, as its name implies, involves the recognition of patterns in all types of data. More simply put, pattern recognition is the search for structure in data [BeP92]. Many times pattern recognition is associated with images; however, patterns exist in other forms of data as well. Two types of pattern recognition exist. The first is called supervised recognition while the second is known as unsupervised recognition

[JDM00]. Supervised recognition is sometimes called learning with a teacher and involves finding a model for correctly associating input data with target data. Unsupervised recognition is often referred to as learning without a teacher and corresponds to cluster analysis and statistics [Jam88]. For the purposes of this research, unsupervised recognition will likely be more useful than supervised recognition because the characterization algorithm will not know what application is accessing the network. While previous methods have explained classification methods for various types of data, pattern recognition takes into account that a measurement for determining classification of some types of data might not exist.

A model that encompasses the elements of a typical pattern recognition system is shown below in Figure 2-6 [BeP92]. This model is referenced throughout the remainder of the discussion on pattern recognition. The resultant model should be able to classify and predict elements of most tested processes.
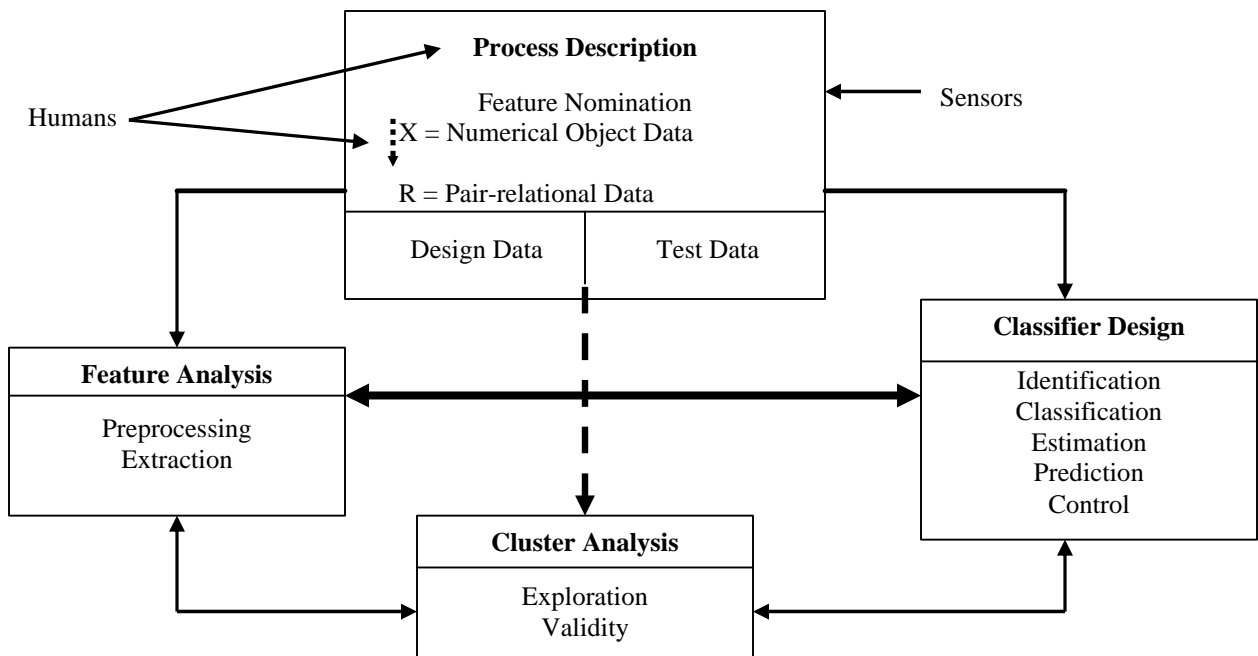
Figure 2-6 Elements Of A Typical Pattern Recognition System

2-15

In addition to the model, the pattern recognition process outlines a series of steps needed

to accomplish the creation of a pattern recognition system

a) Nominate data to be captured.

b) Collect data via simulation and lab tests.

c) Search for underlying structure in the data that provide a basis for hypothesizing

relationships between the variables governing the process.

d) Formalize hypotheses by characterizing the process with equations, rules,

algorithms etc.

e) Propose a model of the system.

f) Analyze various aspects of the model to bring additional insight into the model or

the process it represents.

g) Train the model with labeled training data. This data should be well understood

so that the model is tested, not the data.

h) Test the model with both training data and test data.

i) Build a system that implements the model [BeP92].

Process description is a representation of the process of interest or in other words,

how the process is described. Human and machine inputs can be used to influence

process description. A simple approach is typically used during this phase of pattern

recognition. Numerical object data can be simplified into value-attribute pair data before

an analysis is completed. Once the data is in a form in which it can be analyzed, it can be

grouped into test data and design data. The design data is used for feature analysis while

cluster analysis is performed on the remaining data. Feature analysis is the exploration

and improvement of raw data. During feature analysis, the attributes of the data to be

analyzed are determined. Many different techniques are used to accomplish feature analysis but the oft-quoted line, "keep it simple, but only as simple as it needs to be" is a good rule when considering feature analysis [BeP92].

Cluster analysis is the process of grouping data into subsets where the data in one subset is similar to other data in that same subset yet as different as possible compared to the data in another subset [GrR02]. Cluster analysis is sometimes labeled as fuzzy pattern recognition because several clusters can be formed on a given data set and all can be considered correct. Consider the following example. Let X={father, mother, son, daughter}, a set of four objects. The father has brown eyes while the other three have blue eyes. The mother has type O blood and the others have type A blood. If a cluster analysis was done on the set X, several different correct groups could be formed. For example, if the criterion for a cluster were relationship, the entire set would be included in the group. However, if the criterion for a cluster were eye color, the father would be in one group while the other three would be in a different group. Finally, if the criterion were blood type, the mother would be in one group and the other three would be in a different group. The main point is that cluster analysis can yield several "correct" groups from the same data set. This is why feature analysis is so important. The features considered to be important must be decided before any cluster analysis can begin on the data.

Classifier design is similar to cluster analysis but is more rigorous in labeling data in a set. While cluster analysis tries to label an entire data set with certain attributes, classifier design labels every data point in the data set [BeP92]. As Figure 2-6 indicates, all forms of pattern recognition should benefit from certain aspects of other forms. Each

of the four elements in the pattern recognition system can use each other's results to achieve the greatest potential for an effective pattern recognition system.

2.5   Association

Association algorithms attempt to find correlations among different values in a data set [Str02]. An analogy is that of a supermarket manager attempting to draw an association among the different items in a shopping cart to learn the habits and preferences of a shopper. This same approach can be used when monitoring network traffic and trying to decide whether the traffic is normal or not.

Two stages are involved in associating data. The first step generates item sets with the specified minimum coverage while the second step determines which rules apply to the data to gain the minimum accuracy levels desired. During the first stage, one-item data sets providing at least minimum coverage are generated, and then these one-item sets are used to generate two-item sets. This process is repeated until no more data sets are available that conform to the minimum requirements. This stage is very expensive in terms of time and processing requirement because each combination of data sets is reviewed. This approach may not be optimal, especially for large data sets [WiF00]. The second stage attempts to take the data sets generated in stage one and apply rules that provide minimum coverage. This process is not as time intensive as the first because the search space has been greatly reduced.

The following example will show how item sets are generated during the first phase. Five three-item sets are present for certain network data {ABC}, {ABD}, {ACD}, {ACE}, and {BCD} where feature A, say, is a feature like packet size. The remaining four features could be anything relating to network data. Given that sets

{ABC} and {ABD} have greater than minimum coverage, their union {ABCD} is considered a candidate four-item set. All other sets that can create the four-item set {ABCD} need not be considered because {ABCD} already exists as a potential minimum coverage four-item set. This eliminates the need to consider {ACD} and {BCD} altogether. The only other three-item sets to consider together are {ACD} and {ACE}. The union of these two sets leads to the four-item set {ACDE}. When all possible data sets have been created in phase one, each set is used to generate accuracy rules. These rules must conform to the previously defined levels of accuracy. When all data sets and accuracy rules have been established, an association is built for each data set with the rules. For example, if a final data set included network packet size and packet frequency, a rule might be that the packet size must be medium and the frequency small in order for the network traffic to be considered normal. Association rules are often sought for large datasets, however, the minimum coverage and associated accuracy rules may hinder the optimization of this approach.

2.6 Research Focus

Pattern recognition is the likely candidate for an approach to this thesis research. Some aspects of the various techniques discussed in this chapter may be used in detecting patterns in data. The individual techniques, however, do not provide enough flexibility needed in this research. Pattern recognition includes some aspects of each technique discussed in the chapter. Patterns in the characteristics of network traffic are visually detected and analyzed in order to distinguish among the applications accessing the wireless channel.

2.7 Summary

The focus of this chapter is to provide the background needed to build a methodology for research. The first section discussed some background information on why various techniques are necessary. The next section focused on data mining techniques, namely statistical modeling and naïve Bayesian classifiers, decision trees and covering algorithms. The next discussions considered pattern recognition and association. Finally, pattern recognition was chosen as the primary technique because of its flexibility and robustness.

III. Methodology

3.1   Background

Network packets are organized and built according to various standards. Because of the IEEE 802.11b organization, opportunities exist for monitoring the characteristics of network packets [And98]. This chapter outlines the methodology used to characterize the nature of communication at the application layer of an IEEE 802.11b ad-hoc wireless network.

3.2   Problem Definition

Without knowing the contents of the network packets or the direction of traffic flow, the application accessing the IEEE 802.11b ad-hoc wireless network must be determined. To do this, various characteristics of the packets must be studied. Some of the characteristics which may lead to the characterization of data exchange are packet size, frequency of a given packet size over a given time interval, signal power, inter-packet correlation, and channel utilization. These characteristics are studied and patterns are identified in order to determine the application used in the data exchange. The applications studied include file transfers, web browsing, emails, and printer requests.

3.2.1   Goals and Hypothesis

The research goal is to develop a method for determining what application is accessing an IEEE 802.11b ad-hoc wireless network. This method should determine the application based on certain traffic characteristics. These characteristics are found without knowing the contents of the packet or the direction of the traffic flow. The hypothesis of this research is that certain traffic characteristics can be used to identify the

application accessing an IEEE 802.11b ad-hoc wireless network without knowing the contents of the packet or the direction of the traffic flow.

3.2.2   Approach

The first step is to set up an ad-hoc wireless network. A server machine is configured and used to store web pages, e-mails, file transfers and printers. A second machine is configured to passively listen to the traffic flowing across the network. A third machine is used as a client to access web pages, send e-mails, transfer files, and print documents. With the network correctly configured, traffic is captured and analyzed. The analysis of the traffic should result in the determination of traffic characteristics that allow one application to be distinguished from another. Only network traffic from the IEEE 802.11b standard is studied. The initial data capture is between two computers communicating using this protocol. More computers are added to the network in order to mimic typical network situations. As these additional computers are added, determining the applications being used will likely become more difficult. If the developed algorithm effectively characterizes the data exchange between two computers, one computer at a time is added to the network to test the algorithm in a more complex environment.

3.3   System Boundaries

The term system under test (SUT) refers to the complete set of components that are being purchased or designed for a given study [Jai91]. The SUT in this research is the application identification system used to characterize data exchanges on a network. This system includes a method for accepting network traffic, a method for determining what applications are accessing the network and a method for presenting the output of which

application accessed the network. Things that are not included in the SUT are the type of network, the CPU type and speed, the type of wireless network card used in each machine, and the direction of the network traffic. The component under study (CUS) is a specific component in the SUT whose alternatives are being considered [Jai91]. The CUS in this system is the application identification algorithm used to characterize the applications accessing the network.

3.4   System Services

The service provided by this system is the identification of the application accessing the network. Four outcomes are possible for the service this system provides. The first outcome is successful application identification. Success occurs when the correct application is identified by the system. The second outcome is false positive identification. False positives occur when an application is determined to be used on the network when in fact that application is not used. The third outcome is false negative identification. False negatives are cases where applications are determined as not being used on the network when in fact they are. The fourth outcome is unknown identification. Unknowns occur when a sample of network traffic is captured and analyzed and the application cannot be determined.

3.5   Performance Metrics

The metric used to measure the performance of the SUT is based on a comparison between the known applications accessing the network and the applications identified by the system. For example, if FTP is used to transfer data across the wireless network and the system reports that the data exchange is FTP, the system succeeds. This accuracy

metric is chosen because of the nature of the study. Unlike other systems, this SUT is not concerned with speed but rather with the accuracy of characterizing the data exchange. The system should be able to determine with a certain level of confidence what application is being used to transmit data across the wireless network.

3.6   Parameters

Parameters are defined as system and workload characteristics that affect the performance of the system [Jai91]. System parameters generally do not vary between system installations while workload parameters are a characteristic of users' requests. The next two subsections discuss both the system and workload parameters associated with this study.

3.6.1   System

The system parameters for this study include the protocol for the wireless communication as defined by the wireless network card used in each machine and the method for accepting the input to the system. The protocol of the wireless network is important because the characteristics of network packets with respect to a specific application may change as the protocol changes. The workload accepted into the system must be accurate in order for the pattern recognition algorithm to accurately determine what application accessed the network. If the method for accepting the workload into the system does not correctly accept the workload, the system will likely produce inaccurate results.

3.6.2   Workload

The workload for this study is unknown wireless network traffic. The workload characteristics that affect the performance of the system are the four applications, the size of the time interval used for traffic collection, and the number of computers connected to the wireless network. The vendor and data type used for each application should have no effect on the system. For example, if Microsoft Internet Explorer is used for web browsing on one client and Netscape is used on another client, the system should correctly identify both as HTTP traffic. Also, if one text based e-mail is sent and one image based e-mail is sent, the system should correctly classify both as e-mail. If the size of the time interval increases, more traffic is captured and more patterns are detected in the characteristics of that traffic. When more traffic is available for consideration in the system, more accurate results are possible. If any anomalies exist, they are more likely to be found and recognized when more data is available. This is all accomplished by increasing the length of the time interval for traffic collection. As the number of computers connected to the network increases, the complexity of characterizing data exchange also increases. This could lead to a less accurate description of what application is accessing the network.

3.7   Factors

The three factors chosen for this study are:

1. Number of computers connected to the network

2. Number of vendors used for the applications

3. Type of data used by each application

There are two levels for the number of computers connected to the network. The values are two and three computers. There are four levels for the vendors used for the applications. The value for those levels are 1) Microsoft Outlook Express for e-mail, 2) Microsoft Internet Explorer for HTTP, 3) DOS based command prompt FTP, and 4) remote printing using the Microsoft Office suite to access a shared HP Laser Jet printer. There are three levels for data type. The values for those levels are text, image, and mixed. The number of computers factor is expected to have a high impact on system performance. System performance will likely decrease as the number of computers increases. The number of vendors factor and types of data factor are not expected to significantly influence system performance.

3.8   System Evaluation Technique

The evaluation technique used is a direct measurement of the system. Traffic characteristics associated with a known application are studied. The characteristics of packet size, signal power, and channel utilization are captured as the data is transmitted across the network. The frequency of a given packet size over a given time interval and sequence of packets are determined after the traffic has been captured. For example, if e-mail is transmitted from one computer to another, each packet captured will have its size measured. Figure 3-1 provides a graphical representation of packet size distribution of a sample of e-mail traffic. The packet sizes are placed into a relative frequency histogram to show how often certain packet sizes appear in the time interval used in the data capture.
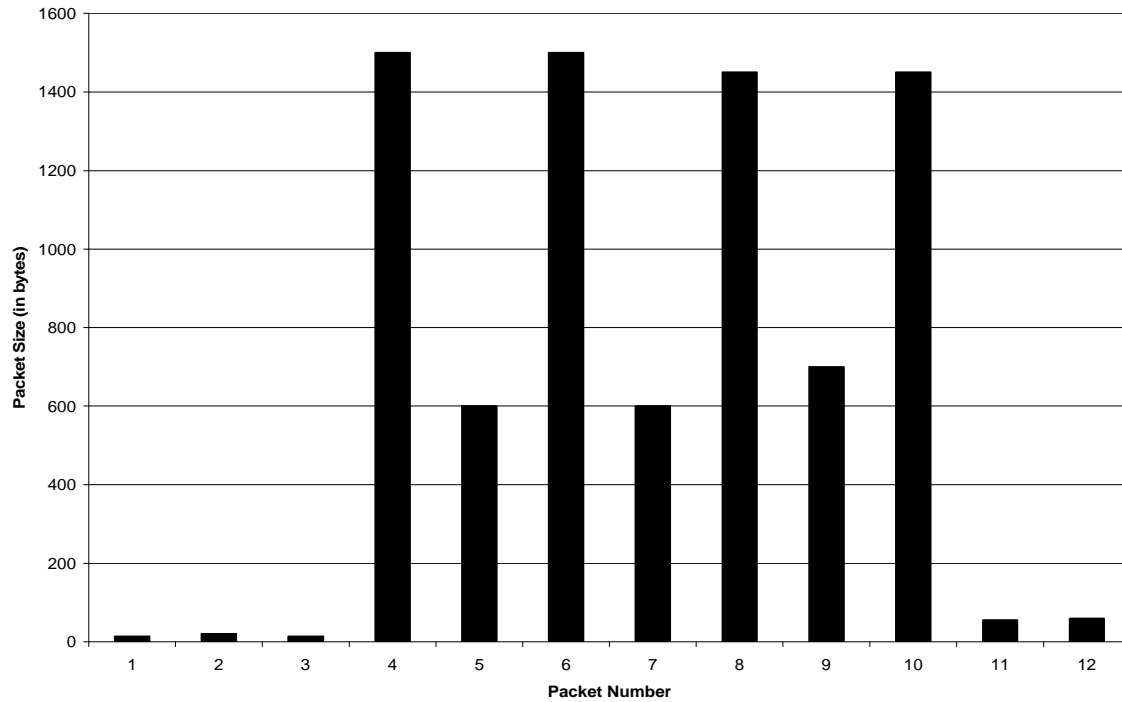
Figure 3-1 Time Series Packet Size Distribution of an e-mail

The diagram in Figure 3-2 shows a histogram of e-mail traffic. The histogram shows that the traffic consists of five small packets, three medium packets, and four large packets. This helps to determine what type of application is used without looking strictly at the shape of the packet size distribution.

Table 3-1 shows all the collected characteristics of the e-mail traffic. The packet size column shows the size (in bytes) of each packet, the packet frequency column shows the number of times a certain packet size appears in the time interval, the signal power column shows whether the signal is strong or weak, and the channel utilization column shows the utilization of the channel being used for the data exchange. The packet frequency column labels packets as small, medium, or large (cf., Figure 3-2 and Table 3-1).
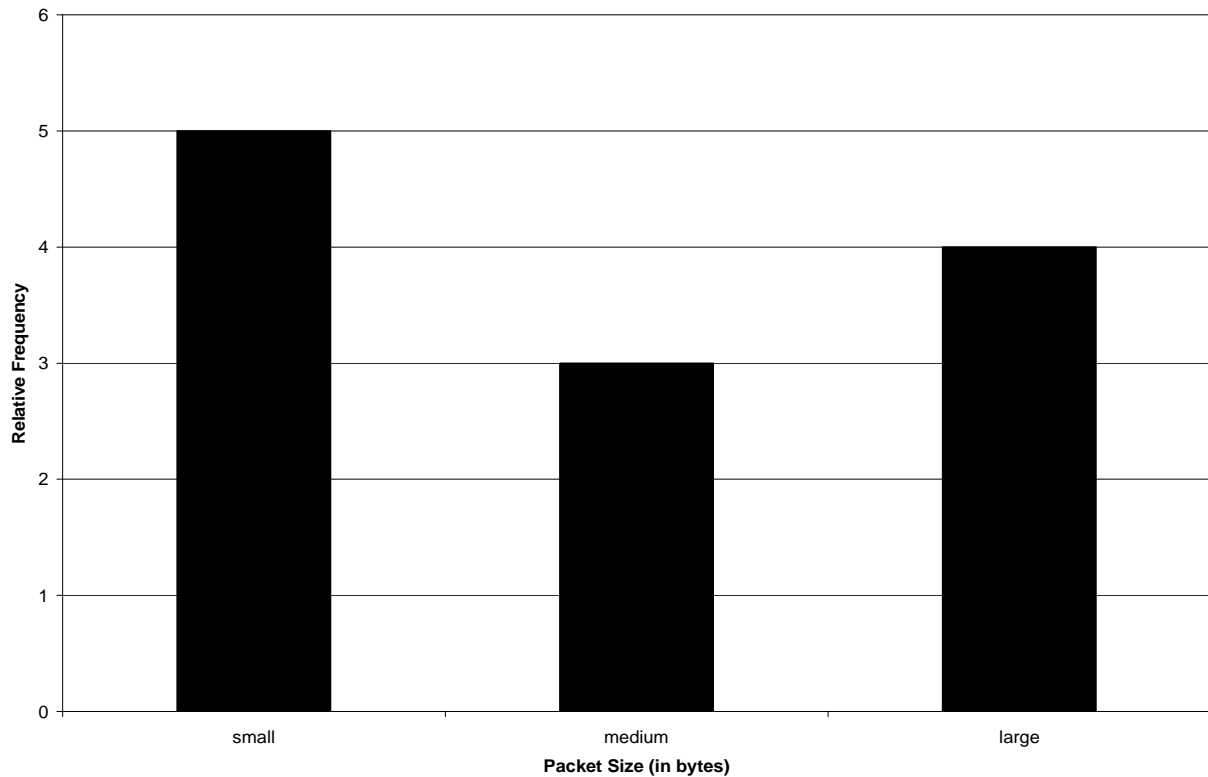
Figure 3-2 Relative Frequency Histogram of e-mail Packet Sizes

Small packets are defined as packets of size 0 to 200 bytes, medium packets are defined as packets of size 201 to 1000 bytes and large packets are defined as packets of size 1001 bytes and larger. Because the application in this example is e-mail, these characteristics can potentially identify future traffic as e-mail and distinguish e-mail from other applications. Other known applications will likely produce different results, and these differences are used to distinguish one application from another. When traffic from all the applications is captured, the characteristics are evaluated and an algorithm describing the differences is established.

Table 3-1 Characteristics of e-mail Traffic

Packet Size	Packet Frequency	Signal Power	Channel Utilization
14	Small (1)	Strong	0.5
20	Small (2)	Strong	0.5
14	Small (3)	Weak	0.5
1500	Large (1)	Strong	0.9
600	Medium (1)	Weak	0.7
1500	Large (2)	Strong	0.9
600	Medium (2)	Strong	0.7
1450	Large (3)	Strong	0.9
700	Medium (3)	Strong	0.8
1450	Large (4)	Strong	0.9
55	Small (4)	Weak	0.5
60	Small (5)	Strong	0.5

3.9 Workload

The workload for this system is IEEE 802.11b ad-hoc wireless network traffic.

The SUT determines what application is used, so the workload must consist of unknown

application traffic. Initially a test workload is submitted with known application traffic of

FTP, HTTP, e-mail, and printer. The true workload consists of network traffic created

using unknown applications. Once an application algorithm is derived, the test workload

is no longer used and the true workload is then used.

3.10 Experimental Design

The response variable for this study is the accuracy of the characterization of data

exchange on an IEEE 802.11b ad-hoc wireless network. The factors affecting this

characterization include number of computers connected to the network, type of data, and

number of vendors for the applications. Because the number of factors is small, a full

factorial design is used for this study. The total number of experiments to run, n, is

$$n = \prod_{i=1}^{k} m_i \qquad \text{(3-1)}$$

$$n = 24$$

where k is the number of factors (3) and m is the number of levels for each factor.

To determine the number of replications to run, the variability of the data is

calculated. In order to decrease the variance of the data, the number of replications for

the experiment is increased. When two replications of each experiment are run, a total of

24 x 2 = 48 observations take place.

3.11 Summary

This chapter outlines the methodology in the study for characterizing the

exchange of data across an IEEE 802.11b wireless network. The first section defines the

problem to be solved and outlines the goal and hypothesis of the study. The system

boundaries and system services are identified next. The performance metrics are listed

and the factors to be changed are drawn from the list of performance metrics. The

evaluation technique and workload are identified in the next two sections. Finally, a full

factorial design is justified because of the small number of factors changing in the study.

IV. Analysis and Findings

The purpose of this chapter is to present and analyze the data collected during the research. A detailed description of the setup and configuration of the network used for testing is found in Appendix A. The following sections present data collected from single and multiple applications. Data collected from two computers transmitting one application at a time is presented first. Data collected from multiple computers transmitting two simultaneous applications is then presented.

4.1 Collected Data

The applications used to generate the workload are File Transfer Protocol (FTP), Hyper Text Transfer Protocol (HTTP), e-mail, and printer. Time series data plots are created to show packet size distributions. These plots reveal distinct patterns that distinguish one application from another without knowing the content of the packets or the direction of the traffic flow. The following subsections provide a detailed description of each application and the characteristics of network traffic that distinguish one application from the other applications.

4.1.1 E-mail

Microsoft Outlook Express version 6 is the vendor used for sending and receiving e-mails. One laptop in the ad-hoc network acts as the Simple Mail Transfer Protocol (SMTP) server and the other client computers send all e-mail to this server (cf., Figures A-1 and A-2). The first set of e-mail traffic contains file sizes between 1 KB and 5 KB. Figure 4-1 shows the time series distribution of the packet sizes for a 1 KB e-mail. This graph shows many small packets (<200 bytes) while only one packet is large.
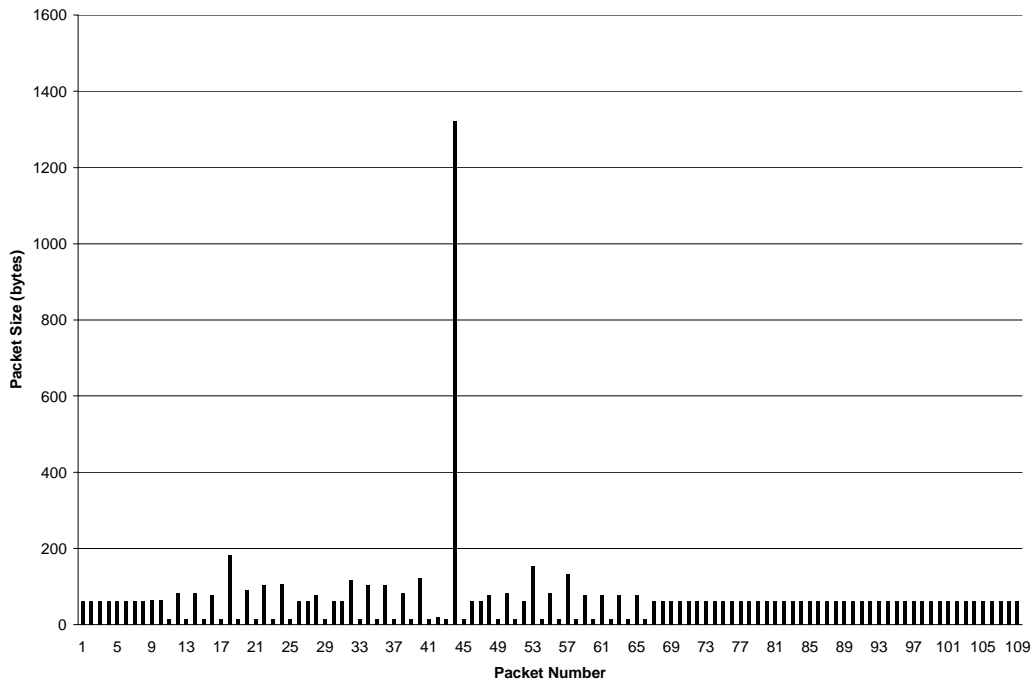
Figure 4-1 Time Series Graph of a 1 KB e-mail

The small packets are protocol packets between the host and client machines. These

packets contain acknowledgement, clear to send, ready to send, and broadcast packets.

The large packet in the middle of this distribution is the data being e-mailed from one

machine to the other. Because the size of the e-mail is small, all the data is placed in one

large packet. As the size of the e-mail grows, more packets are needed to transmit the

data from one machine to another due to the Maximum Transmission Unit (MTU)

restrictions. The default MTU for the wireless cards used in this research is 1500 bytes

[Ent02].

Figure 4-2 shows the time series distribution of the packet sizes for a 5 KB e-mail.

Because the size of the 5KB e-mail is larger than the first e-mail, more than one MTU is

needed to send the data from one machine to the other. In the case of the 5 KB e-mail

(cf., Figure 4-2), three MTUs are needed while one medium sized packet is used to carry the remaining data.
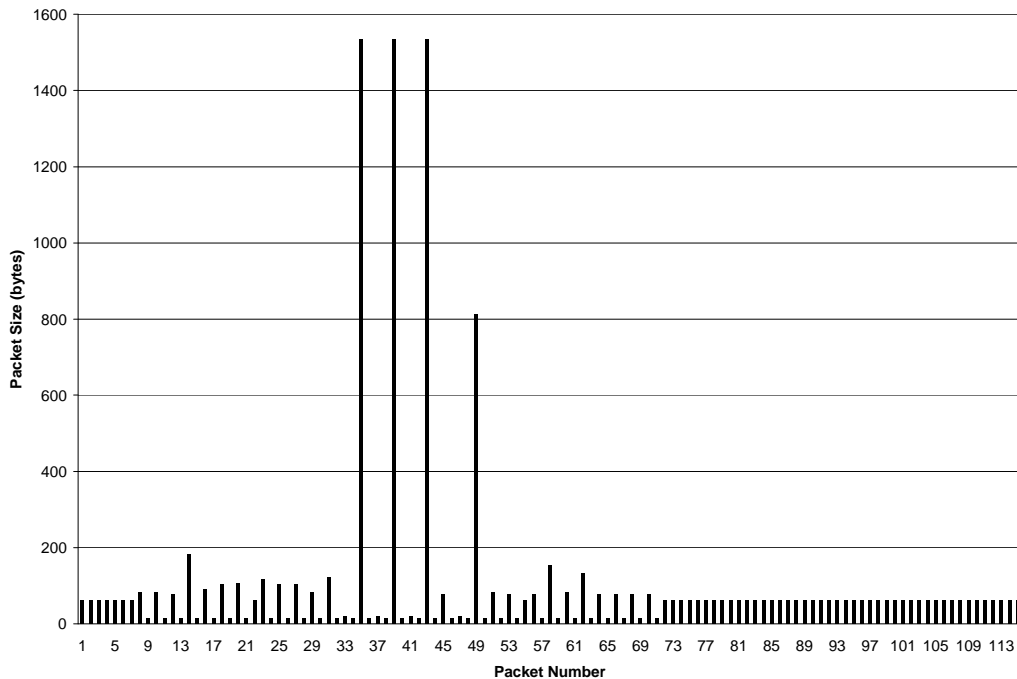
Figure 4-2 Time Series Graph of a 5 KB e-mail

The small packets present in Figure 4-2 are the same type of packets present in Figure 4-1. The only difference between Figure 4-1 and Figure 4-2 is the number of large and medium sized packets. The larger e-mail has more large and medium sized packets.

Figure 4-3 shows the time series distribution of packet sizes from a 55 KB e-mail. This figure shows some interesting characteristics about the distribution of packet sizes of larger e-mails. The small packets that were present in the previous e-mails are also present in this larger e-mail. The difference between the 55 KB e-mail and the smaller e-mails is the number of large and medium sized packets present in the larger e-mail. One would expect several MTUs to be present in a larger e-mail based on the larger amount of data being e-mailed and the size restriction of the MTU.

4-3
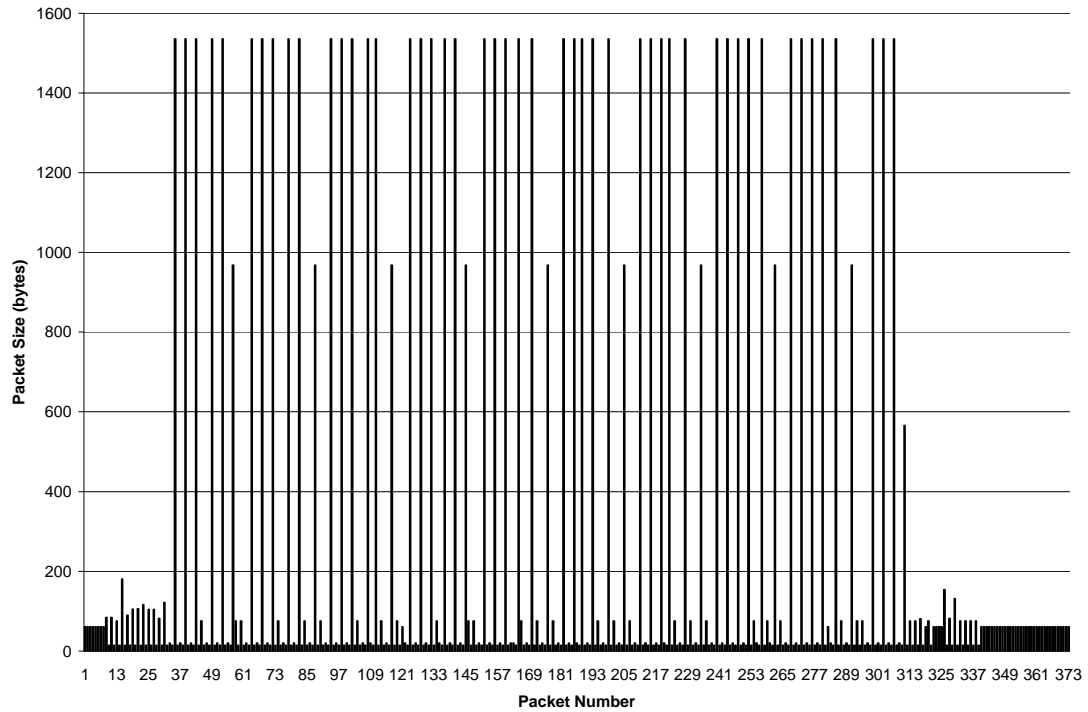
Figure 4-3 Time Series Graph of a 55 KB e-mail

The interesting characteristic shown in Figure 4-3 is the medium sized packets (~950 bytes) that are present throughout the distribution. Figure 4-4 shows the histogram of packet sizes in the 55 KB e-mail.
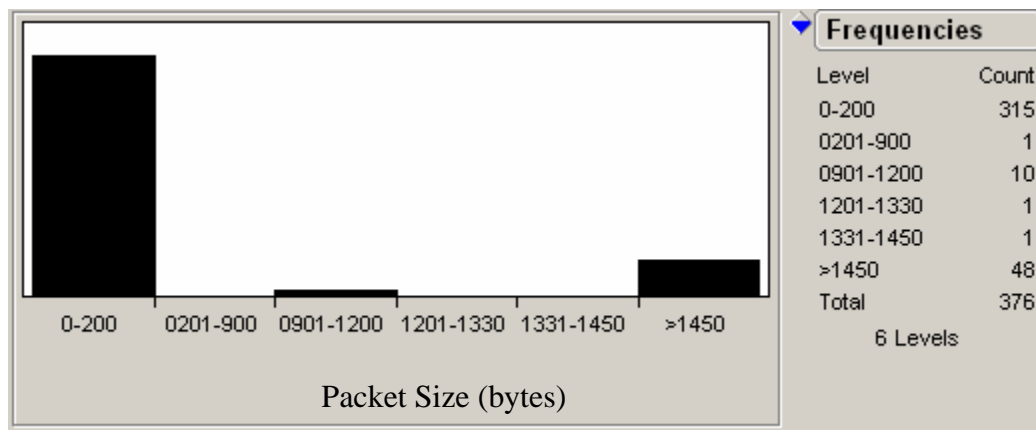
Figure 4-4 Histogram of Packet Sizes from a 55 KB e-mail

This histogram shows that 315 packets of size 0 to 200 bytes are present, 1 packet of size 201 to 900 bytes is present, 10 packets of size 901 to 1200 bytes are present, 1 packet of size 1201 to 1330 bytes is present, 1 packet of size 1331 to 1450 bytes is present and 48 packets of size 1450 bytes and larger are present. Figure 4-5 shows the time series distribution of packet sizes from a 140 KB e-mail.
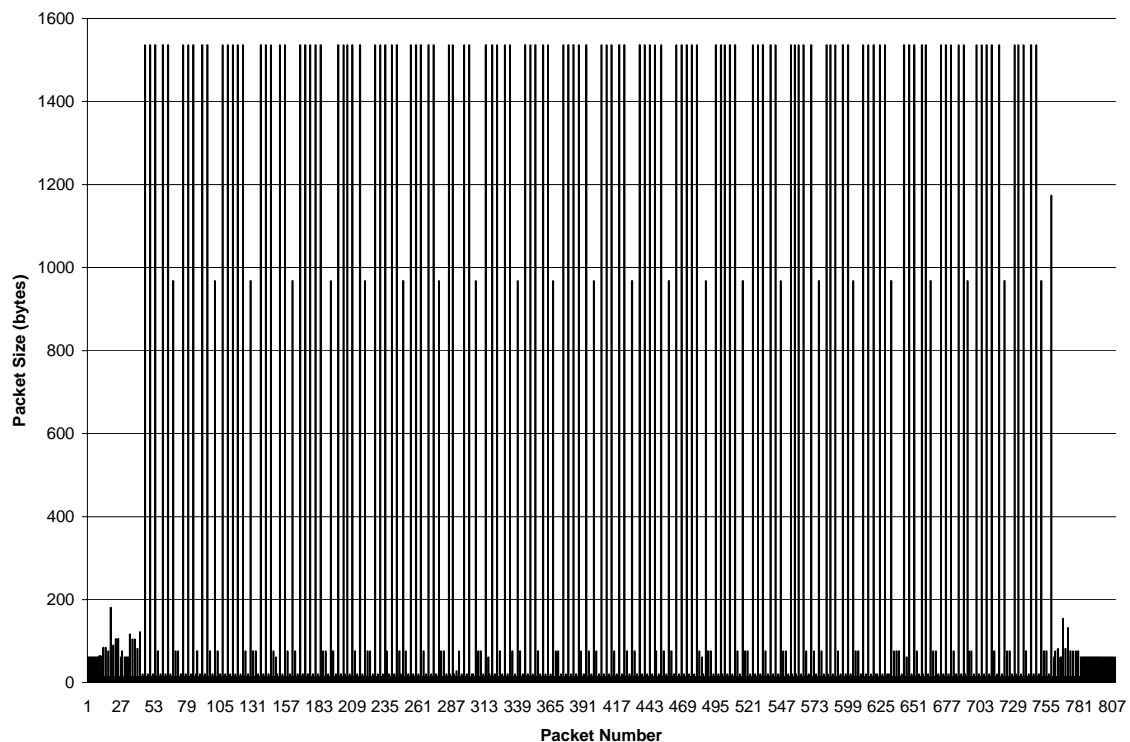
Figure 4-5 Time Series Graph of a 140 KB e-mail

The graph of the 140 KB e-mail also has the small packets that are present in all previous e-mails. As expected, many MTUs are also present in the 140 KB e-mail. The same medium size packets present in the 55 KB e-mail are also present in the 140 KB e-mail suggesting that some sort of pattern is present in the packet size distribution of large e-mails. Figure 4-6 shows the histogram of packet sizes in a 140 KB e-mail.
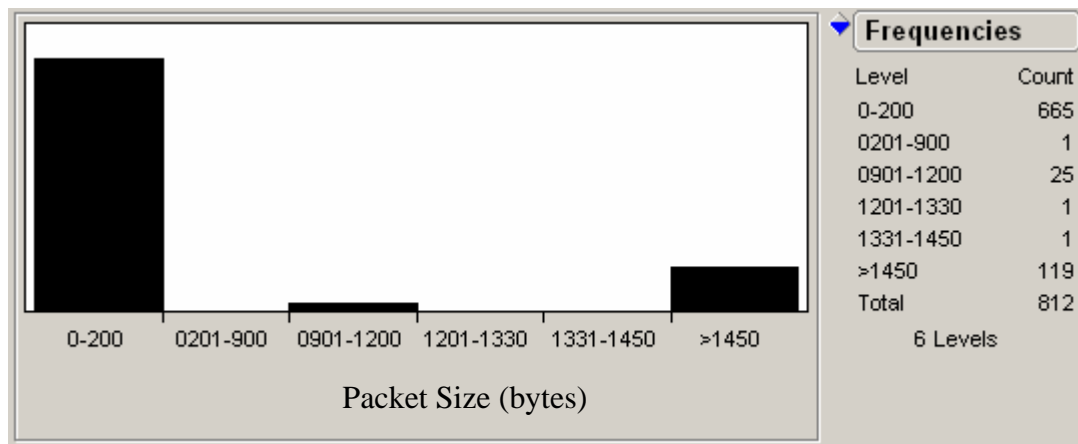
Figure 4-6 Histogram of Packet Sizes from a 140 KB e-mail

Since only one packet each in the groups 201-900, 1201-1330, and 1331-1450 is present,

these packets are not significant in the characteristics defining e-mail traffic.

Interestingly, 25 packets of size 901 to 1200 bytes are present. This is likely due to

protocol layering. When the e-mail data is passed from the application layer to the

transport layer, it is broken up into smaller pieces and delivered to the recipient. If the

application layer provides too much data for the transport layer to transmit at one time,

the transport layer will package that data into several MTUs and the remaining data will

be packaged into a smaller packet. If the transport layer constantly receives the same

amount of data to be segmented from the application layer, a pattern will emerge where

the remaining data will always be placed in the same sized smaller packets. This is what

occurs in e-mails where the file size of the e-mail is larger than approximately 10 KB. In

fact, packets of size 901 to 1200 bytes are only present in e-mails and not in the other

applications. Therefore, when there is a significant number of packets of size 901 to

1200 bytes and an absence of a significant number of packets of size 1201 to 1450 bytes,

the application that likely produced the traffic is e-mail.

4-6

4.1.2 HTTP

The Microsoft Windows 2000 Internet Information Service (IIS) is used to create

and configure an HTTP server. Several different web pages are used. Some of the web

pages are predominantly text, others predominantly image based, and still others are a

mixture of text, image, frames, java scripts, and links. The web pages range in size from

40 KB to 1 MB. Microsoft Internet Explorer version 6.0 and Netscape version 7.0 are

used to access the various web pages. Like the e-mail traffic, HTTP traffic shows unique

characteristics when a web page is large enough for several packet sizes to reach the

MTU. HTTP traffic from small web pages is transmitted using small packets making it

difficult to distinguish it from IEEE 802.11b acknowledgements and broadcasts.

Therefore, traffic from larger (>55 KB) web pages is used.

Figure 4-7 shows the time series distribution of packet sizes from a 140 KB image

based web page accessed by Microsoft Internet Explorer 6.0. The graph of the HTTP

traffic is significantly different than the graph of the e-mail traffic. More small packets

are present in the web page graph. Also, the sizes of the medium packets are larger in the

web page traffic than the medium packets in the e-mail traffic. The large packets are the

same size for both e-mail and HTTP because of the MTU restrictions on packet size.
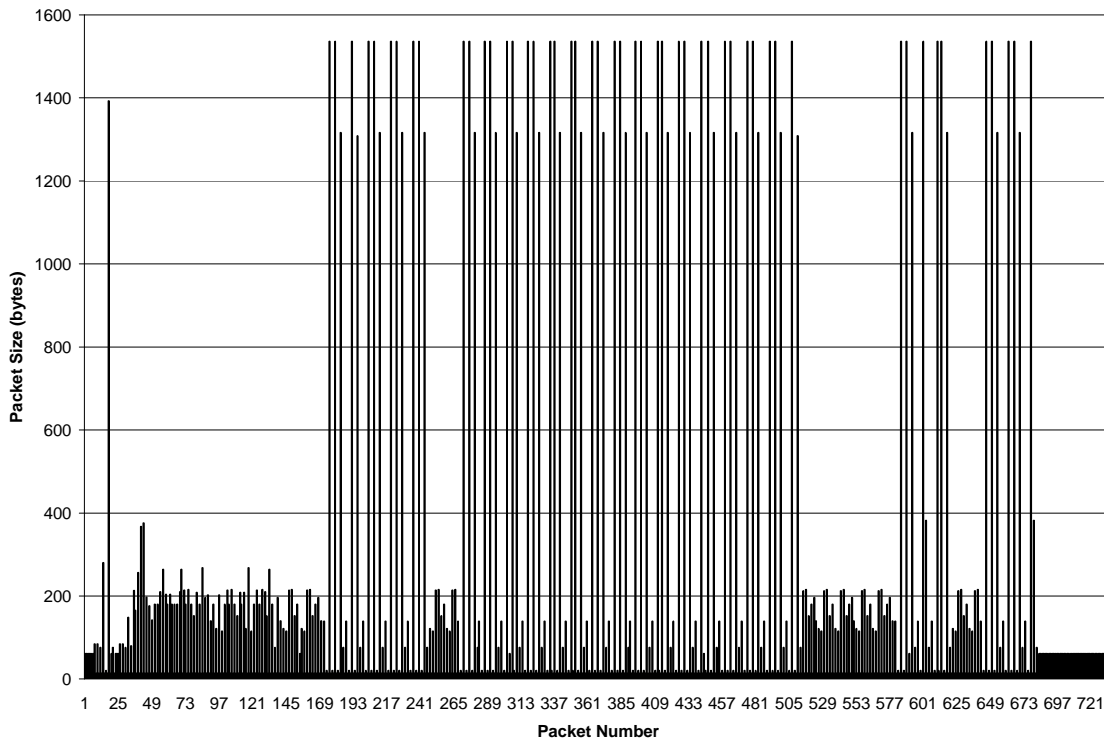
Figure 4-7 Time Series Graph of a 140 KB Image Based Web Page Accessed by
Microsoft Internet Explorer 6.0

Figure 4-8 shows the time series distribution of packet sizes for a 140 KB text based web
page accessed by Microsoft Internet Explorer 6.0. The graph of the text based web page
has fewer small packets than the image based web page. The small packets present at the
beginning of the image based graph are packets exchanged between the HTTP server and
client machines to determine how and where to get the images present in the web page.
Because the text based web page does not have as many images as the image based web
page, there are fewer small packets at the beginning of the graph.

When a web page is accessed by a client machine on the network, the images on
that web page are stored in temporary memory on the client machine. This happens so
the images are quickly available if the same web page is accessed again. For the

purposes of this research, the temporary memory is erased after a web page is accessed. This causes the images to be transmitted across the network each time a web page is accessed.
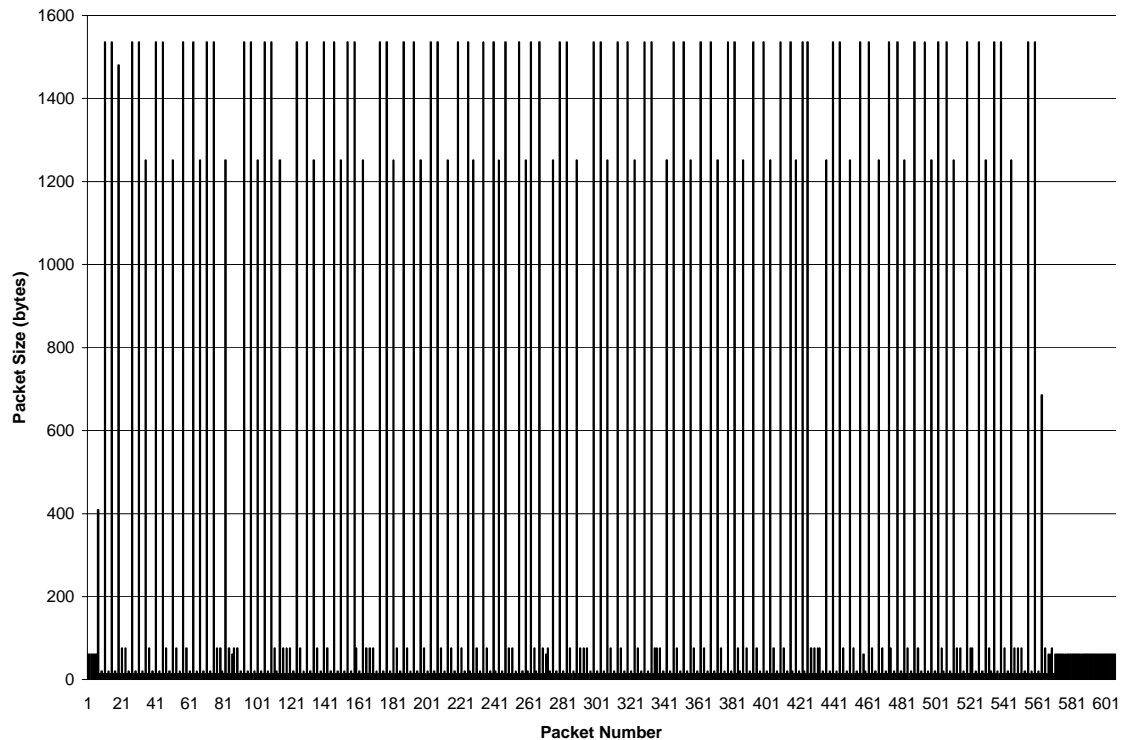
Figure 4-8 Time Series Graph of a 140 KB Text Based Web Page Accessed by Microsoft Internet Explorer 6.0

Figures 4-9 and 4-10 show the histogram of packet sizes in a 140 KB image based web page and a 140 KB text based web page respectively. Both pages are accessed by Microsoft Internet Explorer 6.0. Interestingly, both the image based web page and text based web page have a similar number of packets of size 1201 to 1330 bytes. Also, only one packet of size 901 to 1200 bytes is present in both the image and text based pages.
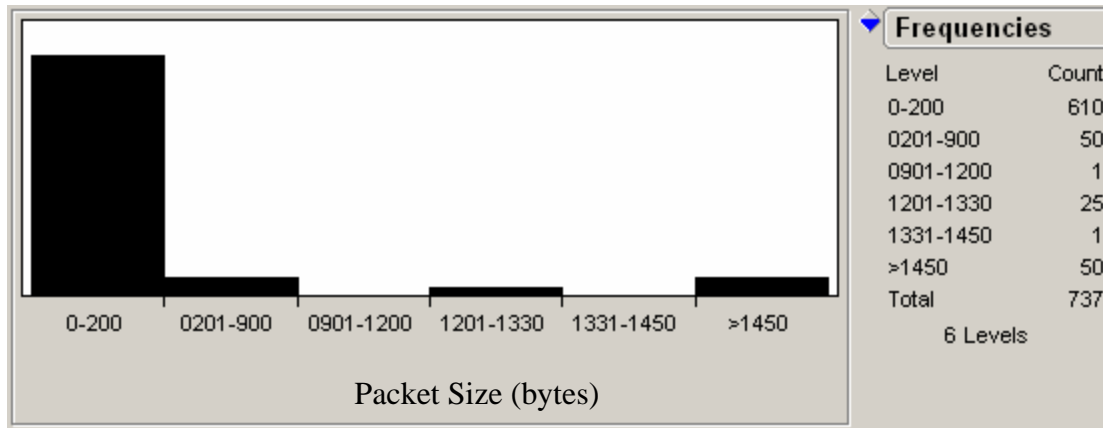
Figure 4-9 Histogram of Packet Sizes from a 140 KB Image Based Web Page

The Frequencies table for Figure 4-9:

Level	Count
0-200	610
0201-900	50
0901-1200	1
1201-1330	25
1331-1450	1
>1450	50
Total	737

6 Levels
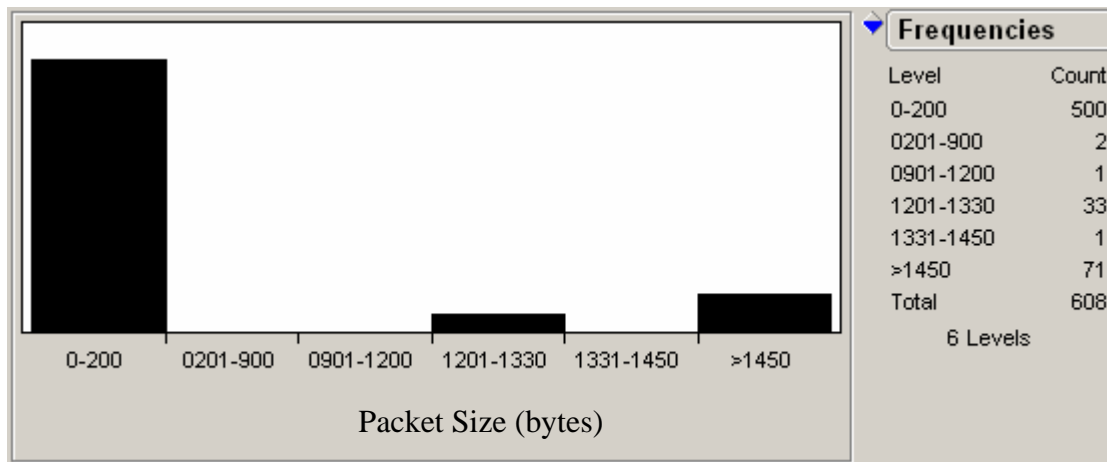
Figure 4-10 Histogram of Packet Sizes from a 140 KB Text Based Web Page

The Frequencies table for Figure 4-10:

Level	Count
0-200	500
0201-900	2
0901-1200	1
1201-1330	33
1331-1450	1
>1450	71
Total	608

6 Levels

This is important because the e-mail data shows that 25 packets of size 901 to 1200 bytes are present where only one packet of size 1200 to 1330 bytes is present for the same size file. The layering of software is, again, the explanation of the smaller packet sizes. It appears that Internet Explorer sends data to the transport layer in different sizes than Outlook Express.

Netscape is another popular tool used to browse web pages. Tests are run using Netscape to determine whether or not Internet Explorer is different than Netscape. Figure 4-11 shows the time series distribution of packet sizes for a 140 KB image based web page accessed by Netscape 7.0.
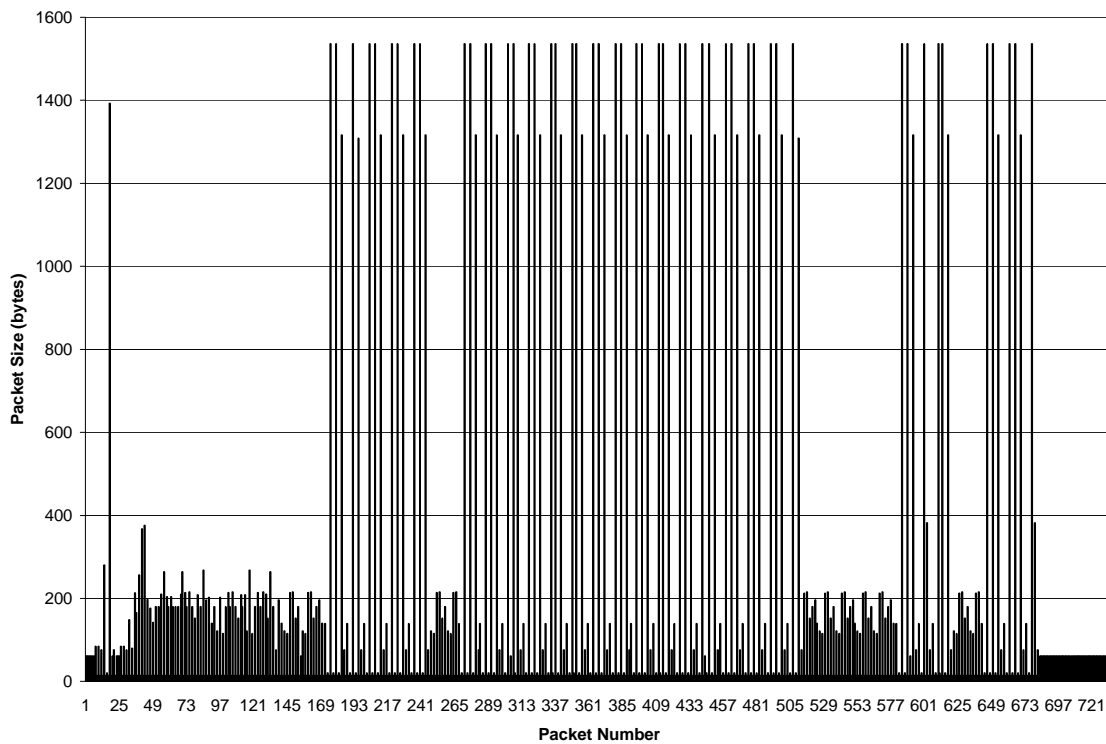
Figure 4-11 Time Series Graph of a 140 KB Image Based Web Page
Accessed by Netscape 7.0

The packet size distribution of the HTTP traffic produced by Netscape appears to be the same as the packet size distribution of the HTTP traffic produced by Internet Explorer (cf., Figure 4-7 and Figure 4-11). In order to completely test the differences between Internet Explorer and Netscape, data from text based web page traffic is collected. Figure 4-12 shows the time series distribution of packet sizes for a 140 KB text based web page accessed by Netscape 7.0. Figures 4-11 and 4-12 have the same packet distribution

characteristics as the respective Microsoft Internet Explorer graphs. So there is no

significant difference between Netscape and Internet Explorer in terms of packet

distribution.
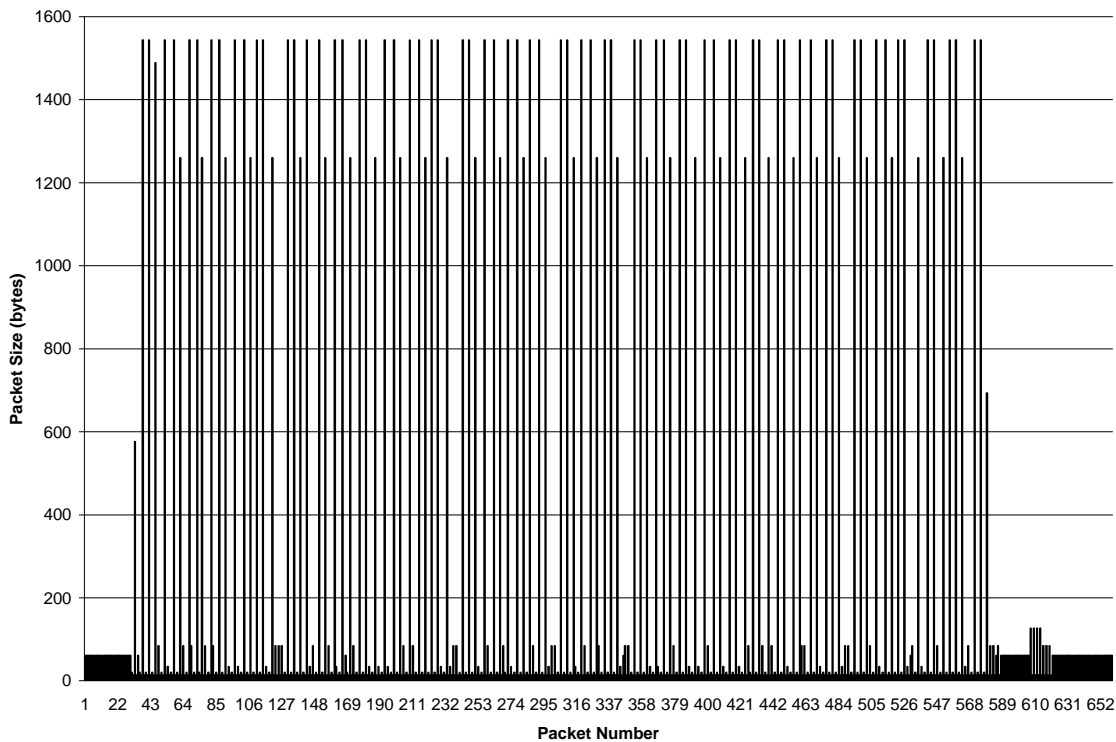
Figure 4-12 Time Series Graph of a 140 KB Text Based Web Page
Accessed by Netscape 7.0

Based on the histograms in Figures 4-9 and 4-10, when there are a significant number of

packets of size 1201 to 1330 bytes and an absence of a significant number of packets of

size 901 to 1200 bytes and 1331 to 1450 bytes, the application accessing the network is

determined to be HTTP. This is true for both Microsoft Internet Explorer and Netscape.

Another unique characteristic in the HTTP traffic is the presence of packets of

size 201 to 900 in a predominantly image based web page and the absence of these sized

packets in a predominantly text based web page. Figures 4-9 and 4-10 are histograms of

HTTP traffic from web pages that are both 140 KB. Figure 4-9 shows packet sizes from a predominantly image based web page while Figure 4-10 has packets from a predominantly text based web page. Figure 4-9 shows a significant number (50) of packets of size 201 to 900 bytes while Figure 4-10 an insignificant number (2) of packets of size 201 to 900 bytes. When packets of size 1201 to 1330 bytes are present in network traffic and packets of size 201 to 900 bytes are also present in that same traffic, the conclusion is made that the traffic is predominantly image based HTTP. When packets of size 1201 to 1330 bytes are present in network traffic and packets of size 201 to 900 bytes are not present in that same traffic, the conclusion is made that the traffic is predominantly text based HTTP.

4.1.3 Printer

The printer configuration is set up using the HTTP and e-mail server machine as the print server. An HP Laser Jet printer is added to the print server and sharing permissions are set so the client machines can remotely print files. Various word processing and spreadsheet files are sent to the printer and the associated network traffic is captured. Files of various sizes are printed and, much like the e-mail and HTTP traffic, unique characteristics of printer traffic appear as the file sizes increase. When small files are printed, characteristics unique to printer traffic are not present. This is because not enough data is present to fill up more than one MTU. File sizes greater than 50 KB are used when capturing the printer traffic to ensure the unique characteristics are present. Figure 4-13 shows the time series distribution of packet sizes for a 55 KB print file. The graph of the printer traffic is different than the graphs of the e-mail and HTTP traffic.
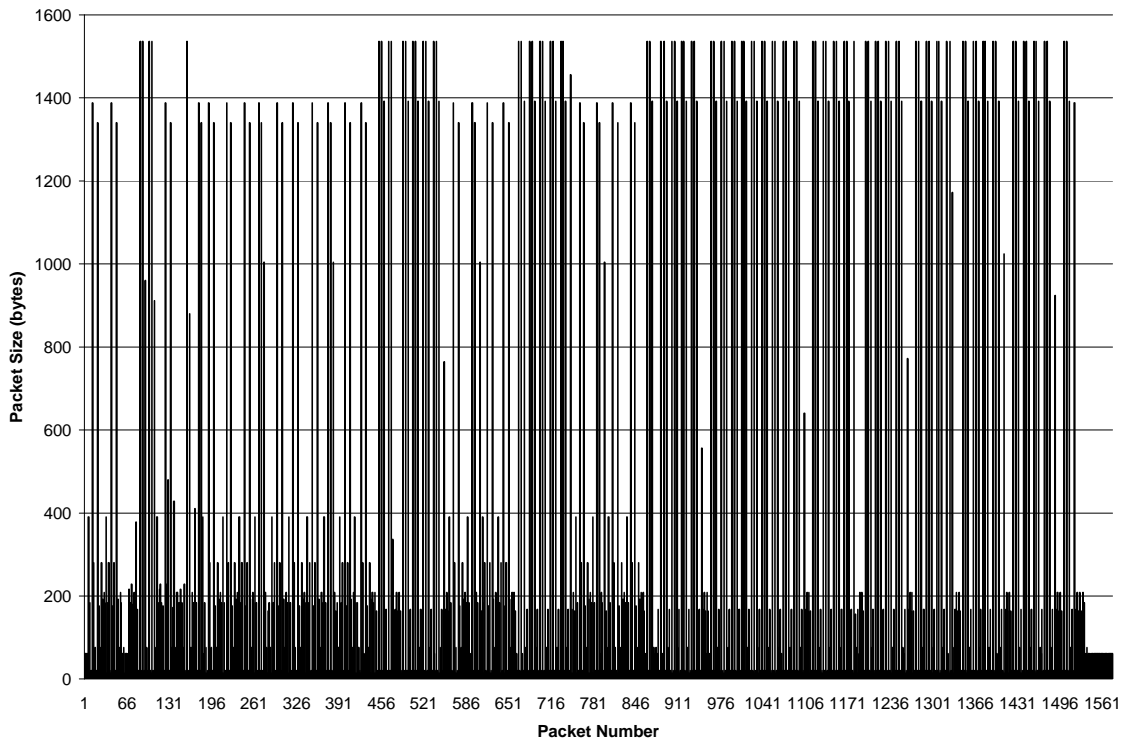
Figure 4-13 Time Series Graph of a 55 KB Printed File

More small packets are present in the printer traffic graph than in the HTTP and e-mail

graphs. Additionally, the sizes of the medium packets are larger in the printer traffic than

the medium packets in the HTTP and e-mail traffic. The large packets are the same size

for all three applications because of the MTU restrictions.

Figure 4-14 shows the grouping of packet sizes in a 55 KB printed file to compare

the packet sizes of printer traffic with e-mail and HTTP traffic. Unlike the e-mail or

HTTP traffic, the printer traffic has a significant number of packets of size 1331 to 1450

bytes. One likely explanation for the difference in the size of the packets in the printer

traffic compared to the e-mail and HTTP traffic is the formatting of data that takes place

before a document is printed.

4-14
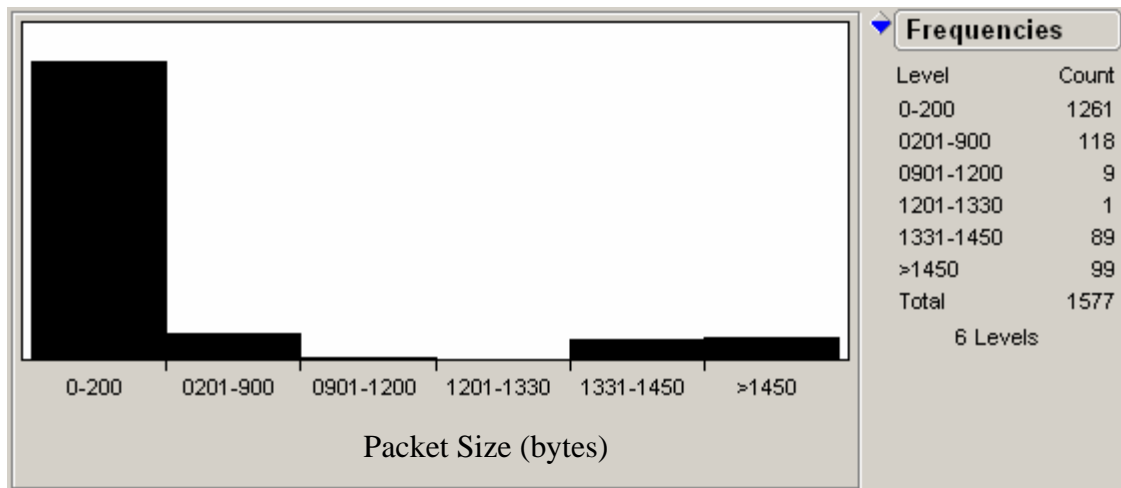
Figure 4-14 Histogram of Packet Sizes from a 55 KB Printed File

This could also account for the difference in the total number of packets in the printer traffic as compared to the other two applications. Approximately 400 packets are needed to send a 55 KB e-mail, 300 packets are needed to access a 55 KB web page, but nearly 1600 packets are used to print a 55 KB file. In order to explore the characteristics of printing larger file sizes, several 140 KB files are printed as well.

Figure 4-15 shows the time series distribution of packet sizes for one of the 140 KB printed files. Although the traffic from the 140 KB printed file has more packets, the distribution of small, medium, and large packet sizes appear to be proportionate to the 55 KB printer traffic shown above. Figure 4-16 shows the histogram of packet sizes in the 140 KB printed file.
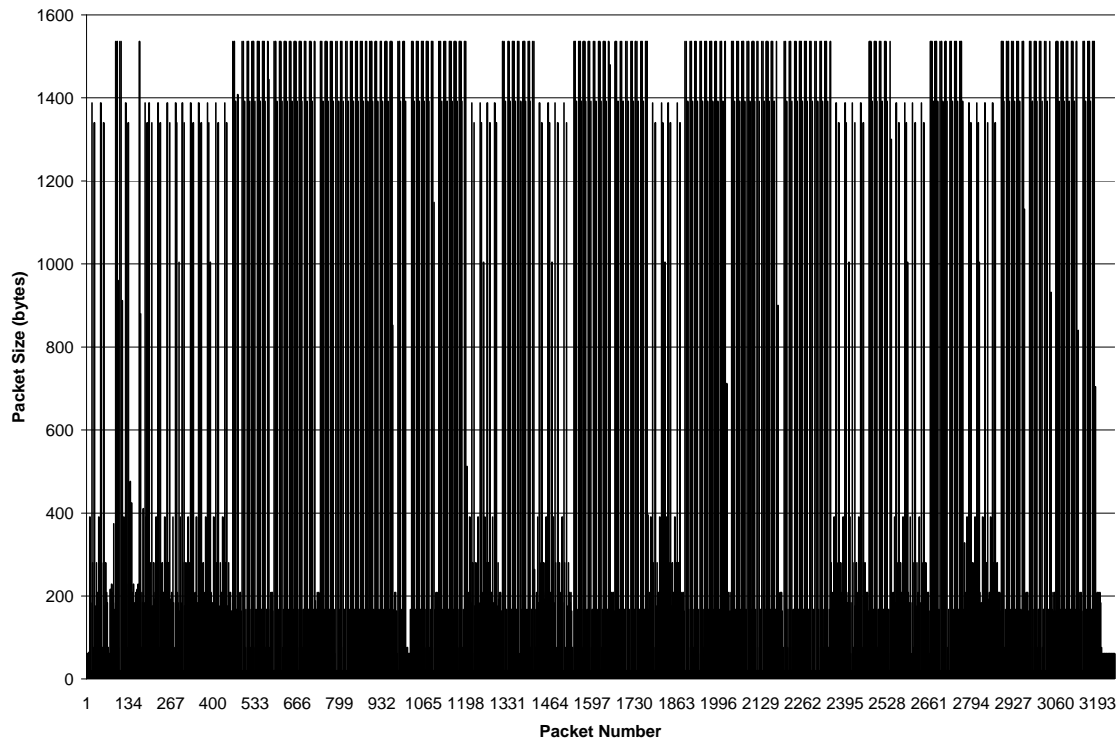
Figure 4-15 Time Series Graph of a 140 KB Printed File
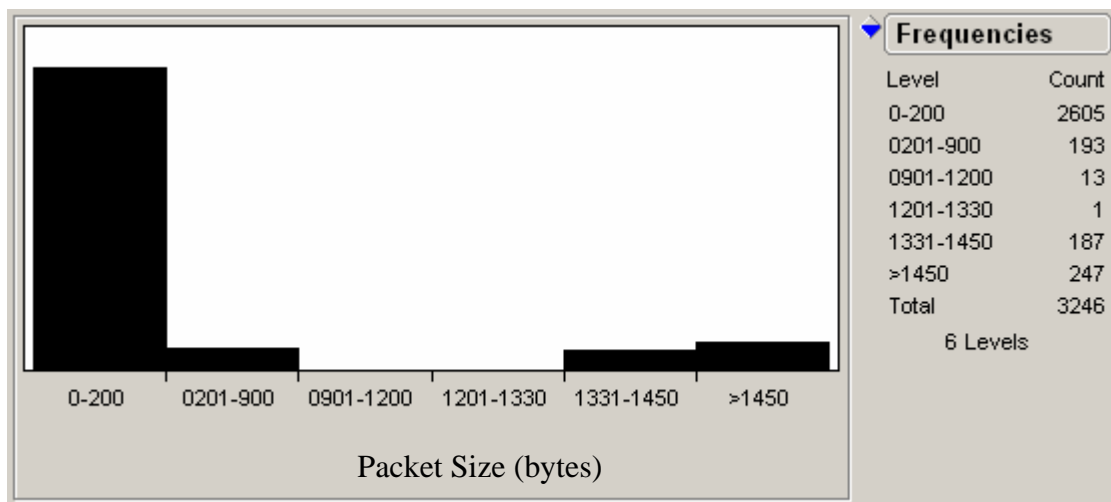
Figure 4-16 Histogram of Packet Sizes from a 140 KB Printed File

When packets of size 1331 to 1450 bytes are present in network traffic, the conclusion is

made that the printer application is used. The packet sizes that made e-mail unique are

packets of size 901 to 1200 bytes, and 13 of these packets are present in this printer

traffic. This could be a problem when trying to identify e-mail versus printer traffic. One

alternative, however, is the fact that the number of packets ranging between 901 and 1200

bytes are fairly insignificant compared to the total number of packets transmitted. In this

example, the packets ranging from 901 to 1200 bytes only comprise 0.4% of the total

traffic while the packets ranging from 1331 to 1450 bytes comprise 5.7% of the total

traffic.

4.1.4   FTP

Microsoft IIS is used to configure the FTP server used for all file transfers. Two

programs are used to test file transfers. The first is a DOS based command prompt

transfer and the second is a Microsoft and Netscape Internet browser based transfer.

Figure 4-17 shows the time series distribution of packet sizes for a 55 KB file transfer

running the DOS FTP program. Unlike the e-mail, HTTP, and printer traffic, the FTP

packet distribution shows significant numbers of small and large packets. The unique

characteristic distinguishing e-mail, HTTP, and printer applications is the size of the

medium packets dispersed throughout the distributions. The distribution of FTP packets

shows the small packets that are present in each application as well as large MTU packets

also present in each application. One could naively assume that the absence of medium

sized packets indicates the presence of the FTP application. However, because

applications other than e-mail, HTTP and printer are present in many networks, other

applications could have time series distributions of packet sizes similar to DOS-based

FTP.
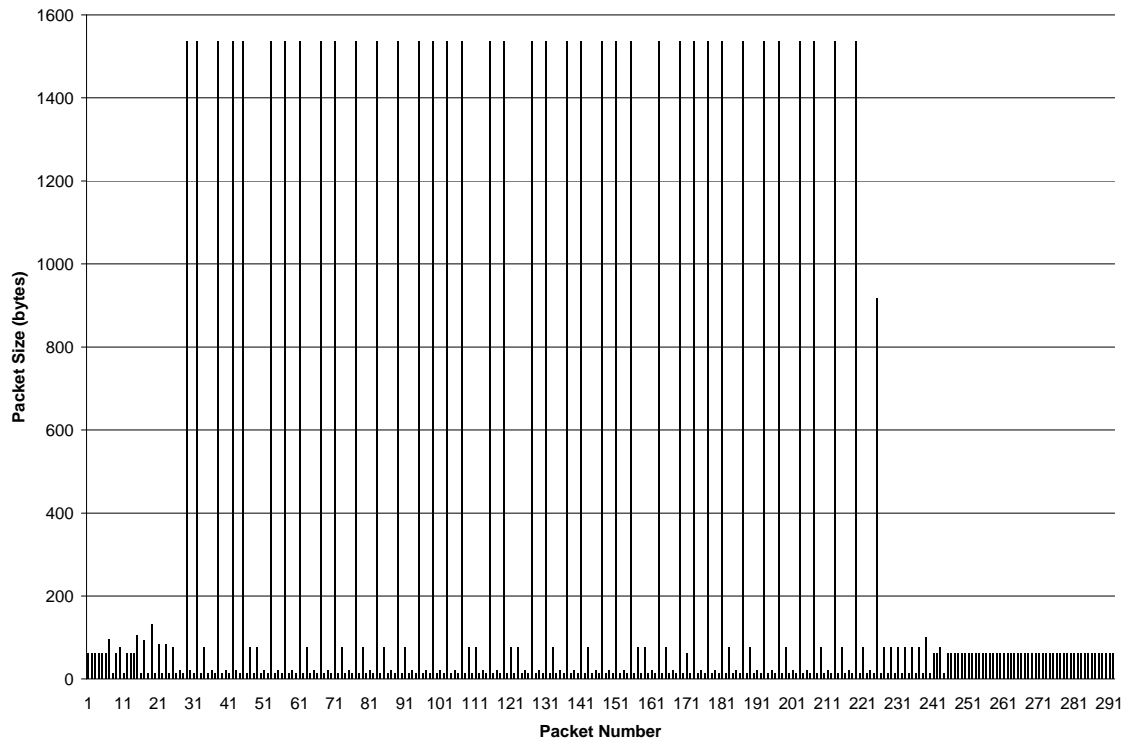
Figure 4-17 Time Series Graph of a DOS Command Prompt 55 KB FTP File Transfer

To ensure that only small and large packets are, in fact, present in the distribution, Figure

4-18 shows the groupings of packet sizes for a 55 KB DOS based file transfer.
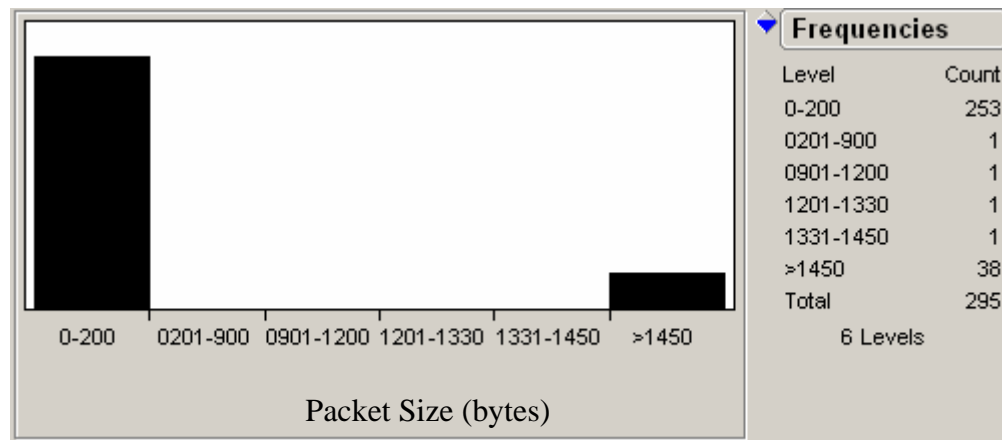
Figure 4-18 Histogram of Packet Sizes from a DOS Command
Prompt 55 KB FTP File Transfer

The histogram above indicates that small and large sized packets are present in the

distribution while insignificant amounts of medium sized packets are present. Larger file

sizes are also tested to ensure the distribution of packet sizes are consistent for all files.

Figure 4-19 shows the time series distribution of packet sizes for a 140 KB file

transfer using a DOS command prompt.
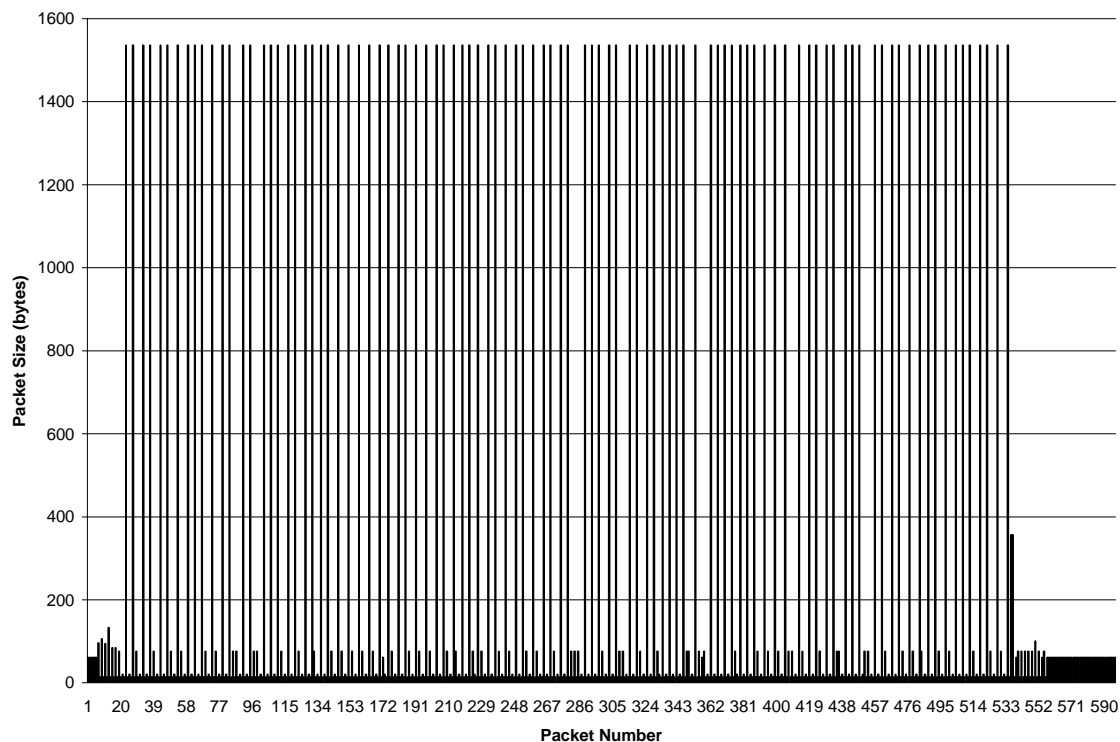
Figure 4-19 Time Series Graph of a DOS Command Prompt 140 KB FTP File Transfer

The graph shows that the distribution of packet sizes from the larger file is proportional

to the distribution of packet sizes from the smaller file size. None of the medium sized

packets that distinguish the other three applications appear to be present in this graph.

The histogram in Figure 4-20 shows the groupings of packet sizes for the same

140 KB DOS based file transfer. This histogram shows that few medium sized packets

are present while a large number of small and large packets are present. This histogram

4-19

is very similar to the 55 KB histogram therefore the larger file sizes do not appear to have

an effect on the shape of the distribution of packet sizes for DOS based file transfers.
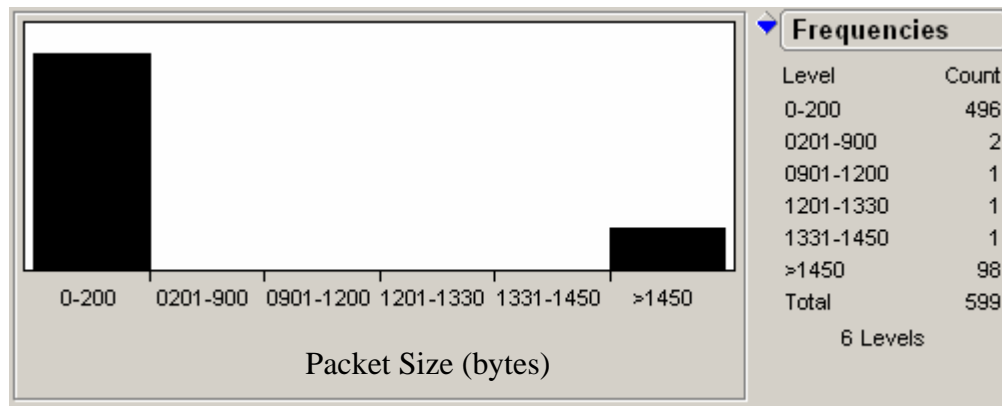
Figure 4-20 Histogram of Packet Sizes from a DOS Command
Prompt 140 KB FTP File Transfer

Microsoft Internet Explorer and Netscape are also used to transfer files. Figure 4-

21 shows the time series distribution of packet sizes for a 140 KB file transfer using

Microsoft Internet Explorer. The packet size distribution of the Internet browser based

FTP looks the same as the packet size distribution for text based HTTP traffic accessed

by Internet Explorer or Netscape. These results could prove problematic when trying to

distinguish between HTTP and FTP applications. The image based web pages show

significant numbers of packets of size 201 to 900 bytes and 1201 to 1330 bytes while the

text based web pages show significant numbers of packets of size 1201 to 1330 bytes.

Figure 4-22 shows the groupings of packet sizes for the 140 KB Internet browser based
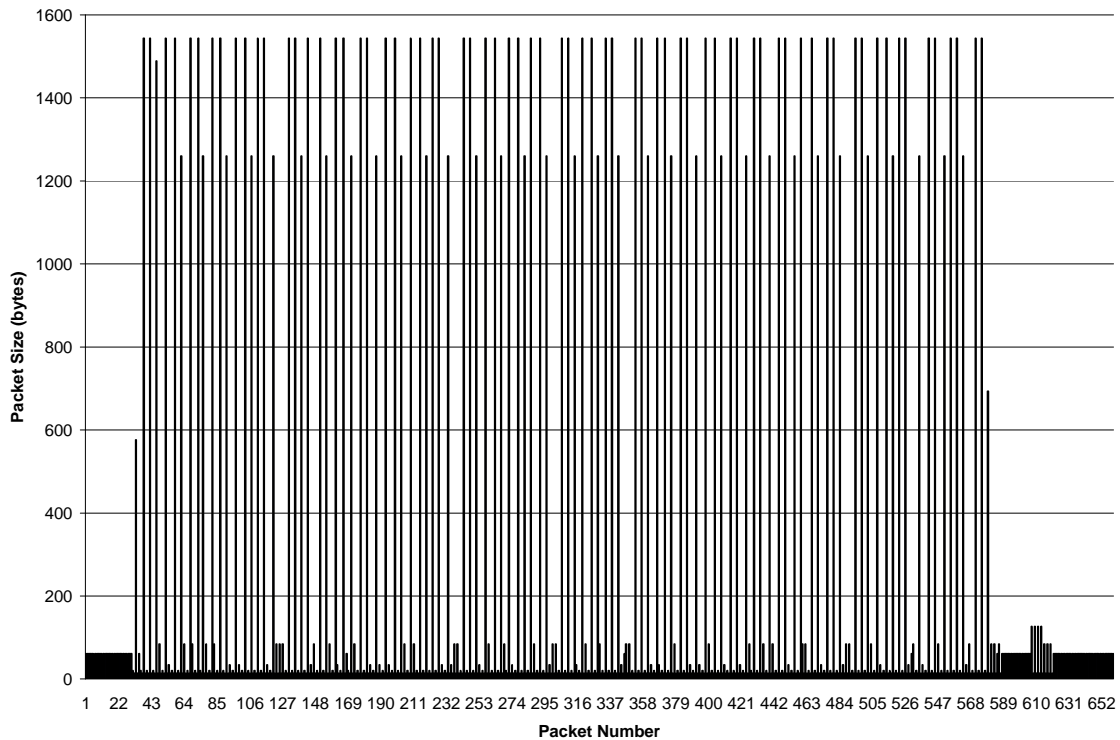
file transfer.

Figure 4-21 Time Series Graph of an Internet Browser Based 140 KB FTP File Transfer

Frequencies	
Level	Count
0-200	553
0201-900	2
0901-1200	1
1201-1330	33
1331-1450	1
>1450	71
Total	661
6 Levels	
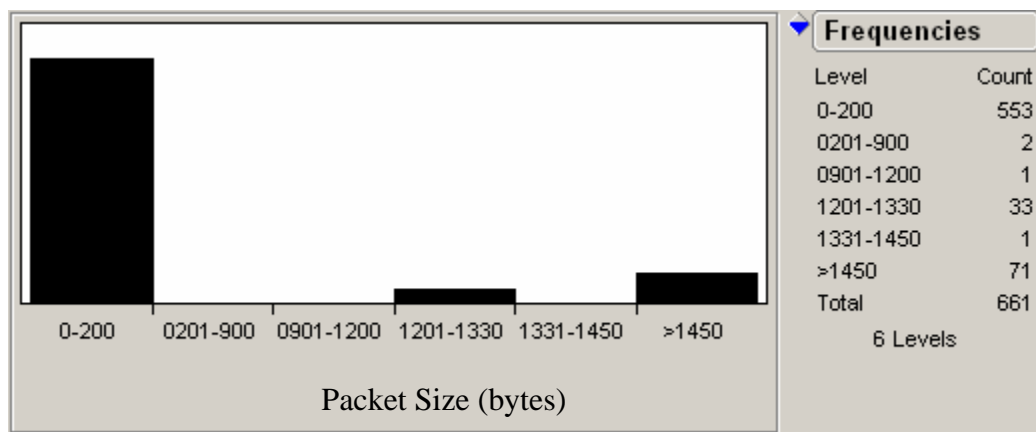
Figure 4-22 Histogram of Packet Sizes from an Internet Browser
Based 140 KB FTP File Transfer

This histogram reveals that a significant number of packets of size 1201 to 1330 bytes are

present in the FTP traffic, but an insignificant number of packets of size 201 to 900 bytes

are present. From this histogram, the conclusion is made that the application being used

is either HTTP accessing a predominantly text based web page or FTP using an Internet

browser to transfer files. Several tests are conducted on file transfers using an Internet

Browser. The results show that the packet size distribution of Internet browser based file

transfers do not change according to the type of data being transferred. Therefore, an

Internet browser based image file transfer looks the same as an Internet browser based

text file transfer. This is important because the characteristics distinguishing image based

web page HTTP traffic are still unique. The FTP results support the conclusion that

when packets of size 1201 to 1330 are present, an Internet browser is used to either

transfer files or access web pages.

4.1.5   Multiple Applications

More than one application is usually being used at any given time in wireless

networks. For this reason, after the single application data is collected from the wireless

network, another client machine is added to the network to capture traffic from multiple

applications. Based on the results shown in the previous histograms, the addition of

multiple applications should have little effect on determining what application is

accessing the network. The first tests using multiple applications involve two client

machines sending e-mail simultaneously. Figure 4-23 shows the time series distribution

of packet sizes for several e-mails. This graph shows that many small packets as well as

many MTU packets are present in the multiple e-mail traffic. Conveniently, the key

characteristic medium sized packets are also present in this traffic.
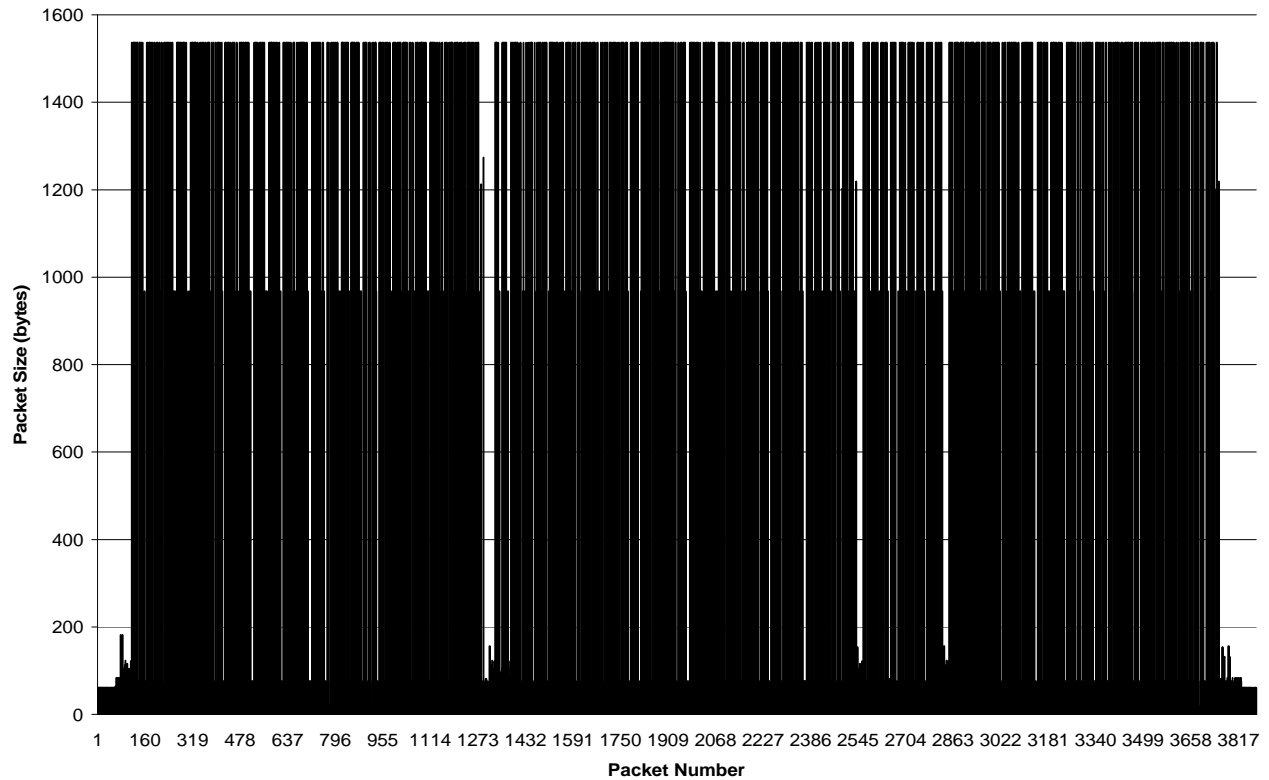
Figure 4-23 Time Series Graph of Several e-mails

These medium sized packets are the exact same size as the medium sized packets present

in the single application e-mail traffic. Figure 4-24 shows the groupings of packet sizes

for the multiple e-mails.

Frequencies	
Level	Count
0-200	3003
0201-900	1
0901-1200	144
1201-1330	6
1331-1450	1
>1450	723
Total	3878
6 Levels	
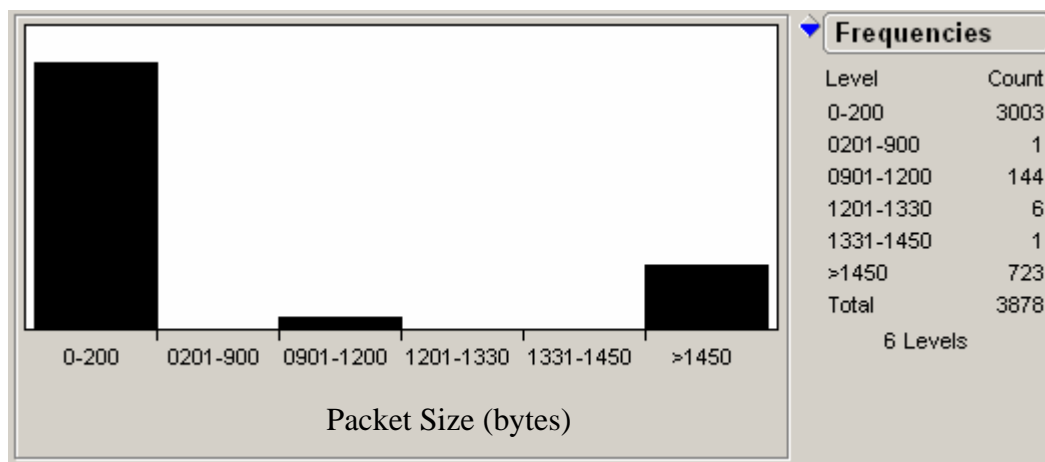
Figure 4-24 Histogram of Packet Sizes from Several e-mails

This histogram shows that a significant number of packets of size 901 to 1200 are present while insignificant numbers of packets of size 1201 to 1450 are present. Because of this, the conclusion is drawn that the application that likely produced the traffic was e-mail. Note that the number of e-mails and the size of each e-mail are not determined by this histogram. The fact that e-mail is the application accessing the network is the only conclusion drawn.

After conducting tests using multiple e-mails, another series of tests using e-mail and HTTP is conducted. The addition of HTTP should not affect the accuracy of the determination of the application because the packet sizes should still follow the same distribution based on software layering. Figure 4-25 shows the time series distribution of packet sizes for e-mail and HTTP traffic.
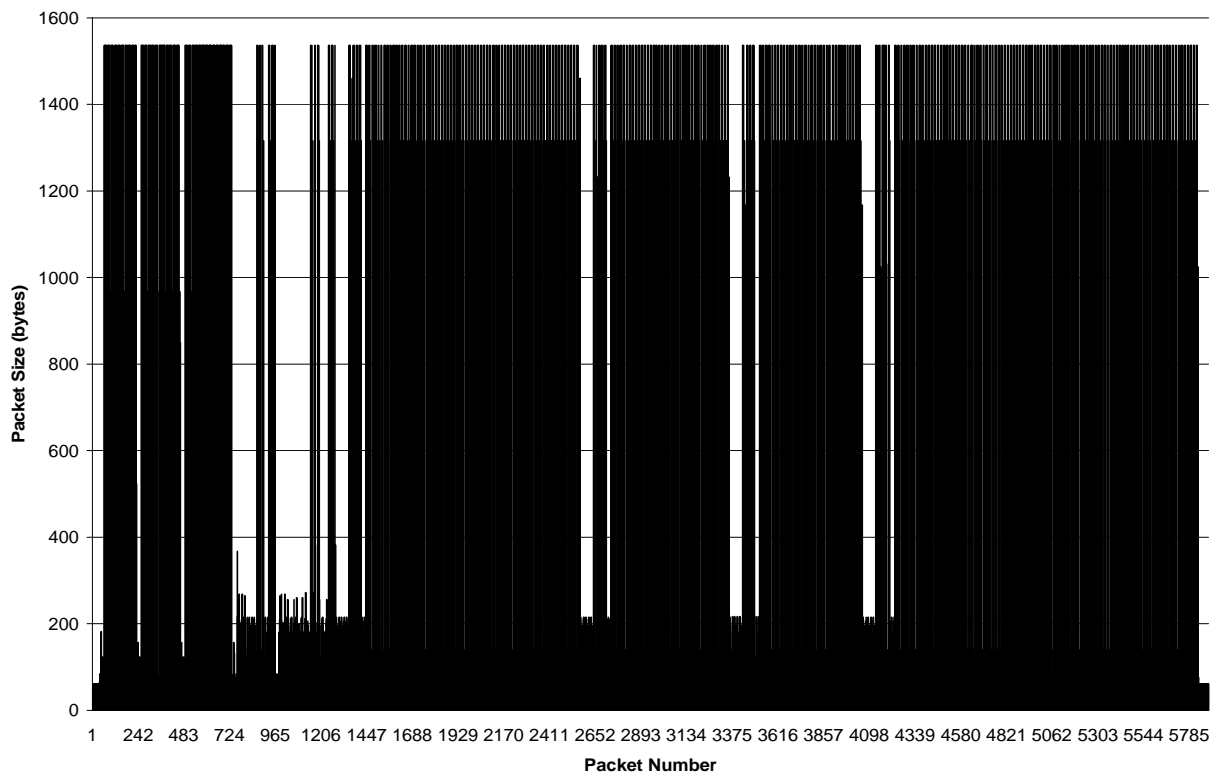
Figure 4-25 Time Series Graph of Several e-mails and Web Pages

4-24

The web pages accessed during this test are image based, text based, and a combination of the two. The small packets between 724 and 1206 on the x-axis (cf., Figure 4-25) appear to be characteristic of an image based web page. The histogram in Figure 4-26 shows the grouping of packet sizes for the e-mail and HTTP traffic shown above.
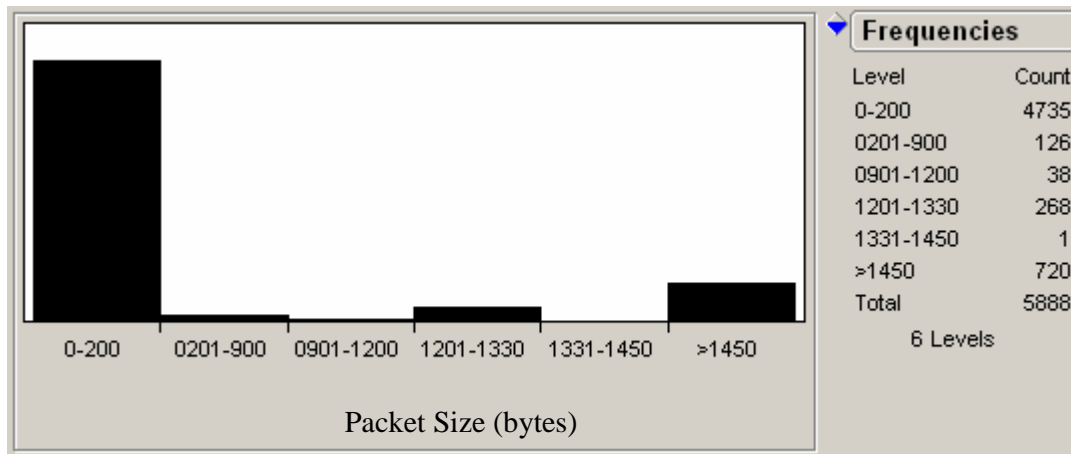
Figure 4-26 Histogram of Packet Sizes from Several e-mails and Web Pages

Clearly, the small packets as well as the large MTU packets are present in this traffic stream. The fact that packets of size 201 to 900 bytes, along with packets of size 1201 to 1330 bytes, are present indicates that HTTP is one application accessing the network, and the web page being accessed contains at least some images. The FTP application could have also been used because of the presence of packets of size 1201 to 1330 bytes. Because a significant number of packets of size 901 to 1200 bytes are present, e-mail is probably also present in this traffic. Figure 4-27 shows the time series distribution of packet sizes for simultaneous e-mail and printer traffic.
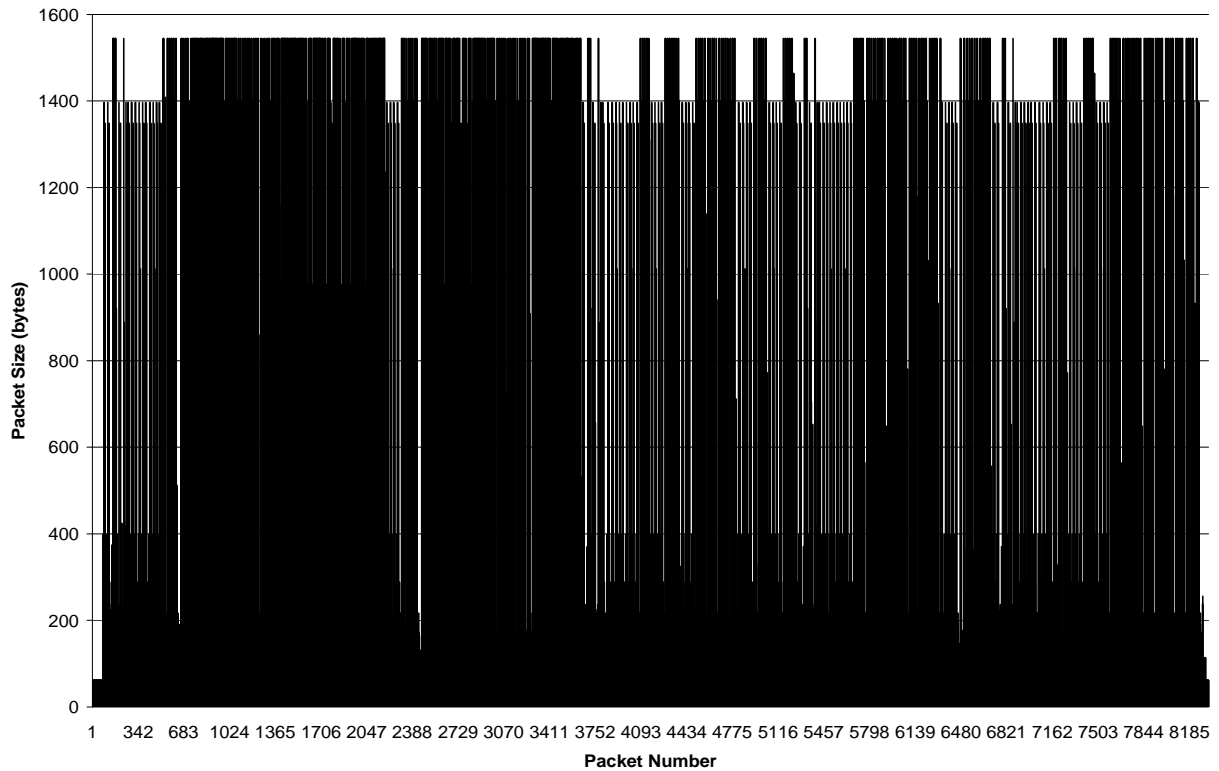
Figure 4-27 Time Series Graph of Several e-mails and Printed Files

The graph reveals that many small packets are present along with large MTU packets.

Conveniently, the medium sized packets that characterize e-mail and printer applications

are also present. These medium sized packets appear to be the same size whether another

application is present in the traffic stream or not. Because of this, the following

histogram in Figure 4-28 reveals the various applications being used on the network. The

histogram reveals that a significant number of packets of size 901 to 1200 bytes are

present. This indicates the presence of the e-mail application in the traffic stream. Also,

many packets of size 1331 to 1450 are present indicating the use of the printer application

as well. Many packets of size 201 to 900 bytes are present, but the insignificant number

of packets of size 1201 to 1330 bytes leads to the conclusion that these smaller packets

are meaningless in determining what type of application is being used.
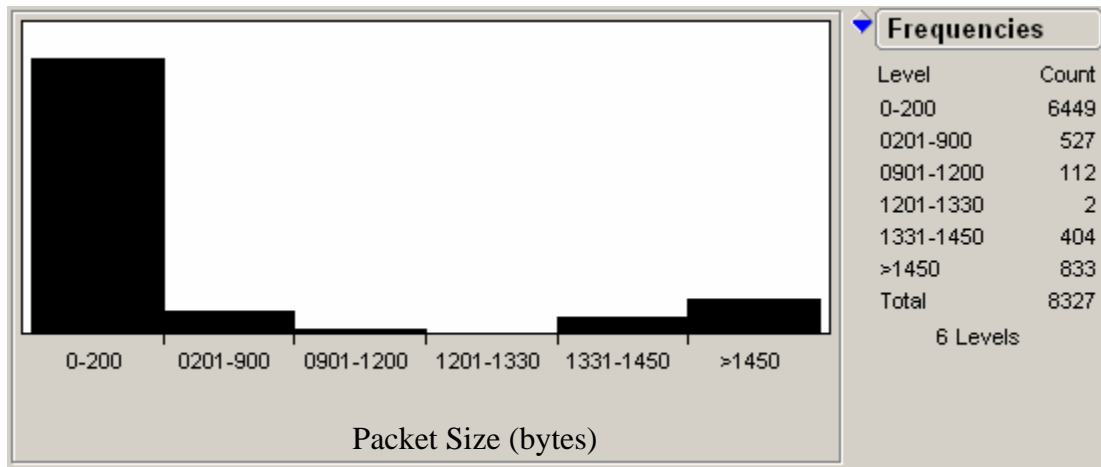
Figure 4-28 Histogram of Packet Sizes from Several e-mails and Printed Files

Figure 4-29 shows the time series distribution of packet sizes for various web pages being accessed by Microsoft Internet Explorer and Netscape.
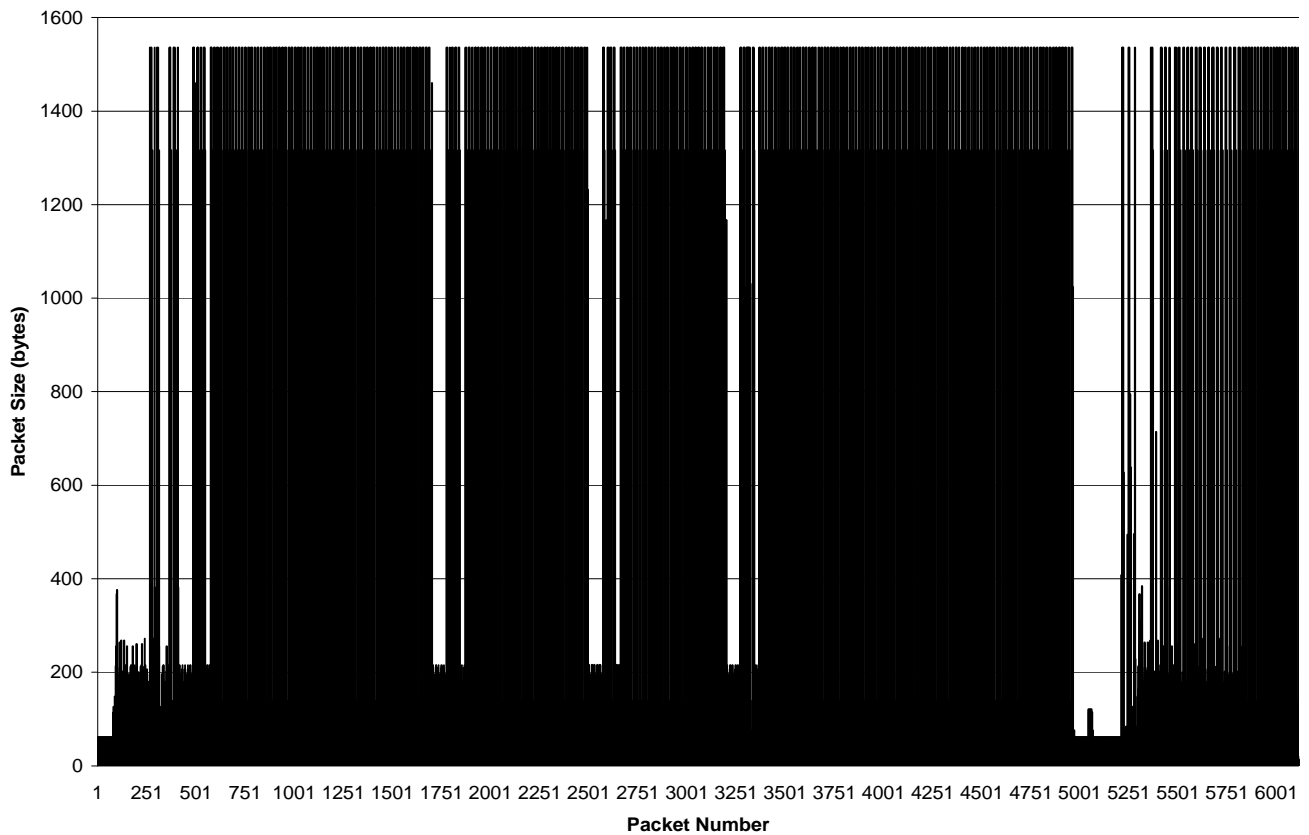
Figure 4-29 Time Series Graph of Several Web Pages Accessed by
Internet Explorer and Netscape

4-27

This graph shows small packets along with the large MTU packets are present regardless of the application. Several medium sized packets are also present, and these packets reveal the application being used in this traffic stream. Figure 4-30 shows the histogram of packet sizes for the HTTP traffic shown above.

Frequencies	
Level	Count
0-200	15402
0201-900	300
0901-1200	15
1201-1330	1024
1331-1450	1
>1450	2082
Total	18824
6 Levels	
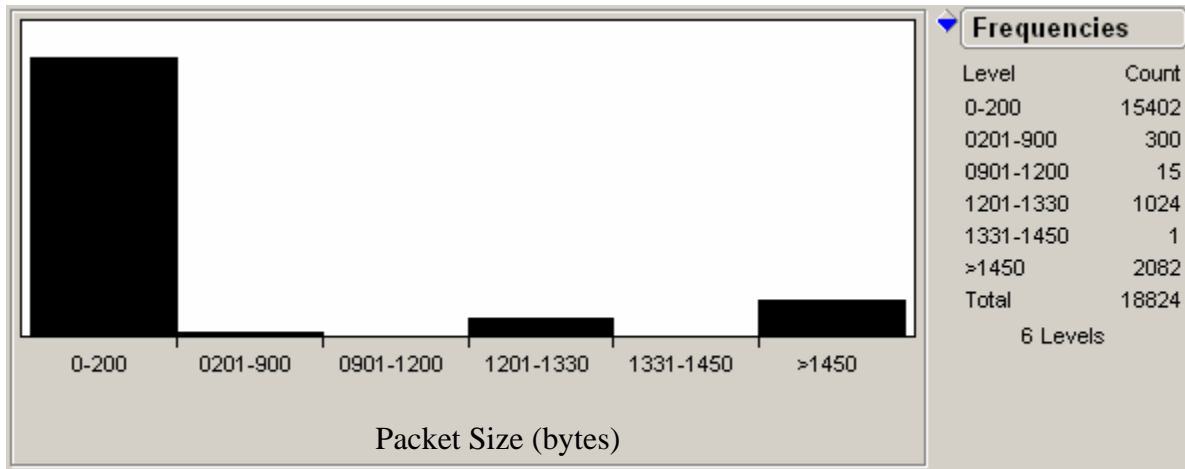
Figure 4-30 Histogram of Packet Sizes from Several Web Pages Accessed by
Internet Explorer and Netscape

The histogram shows that a significant number of packets of size 1201 to 1330 bytes are present. Based on previous discussions, these packets indicate that HTTP or FTP is the application being used. Because packets of size 1201 to 1330 are present, the presence of packets of size 201 to 900 bytes reveals that HTTP is used and at least some images are present in the accessed web pages. Fifteen packets of size 901 to 1200 are also present in this traffic stream. This could lead to a false conclusion that a small e-mail is also present in this stream.

Although more combinations of applications could have been tested, the results from these previous tests suggest that the applications are determined based on the

histograms regardless of the combination of applications. For this reason, not all combinations of applications are tested.

4.1.6   Utilization

Along with packet size graphs and histograms, the average utilization is also explored as a potential characteristic to determine what type of application is being used in a given traffic stream. The channel capacity for the Enterasys IEEE 802.11b wireless card used for each application is 11 Mbps. Therefore, 11,000 Kilo bits per second reveal a fully utilized channel. Table 4-1 shows the average utilization in Kilo bits per second for a given file size and a given application. The time intervals used to capture the various files are not identical for all applications. For example, the 55 KB e-mail transmission did not take as long as the 55 KB print transmission, so the averages were computed over different time intervals. The purpose of the utilization study is to capture the entire transmission of each application. This is done in order to determine the channel utilization of an entire transmission of a given application.

Table 4-1 Average Utilization (Kbits/sec)

	55 KB File	140 KB File
e-mail	157.3	395.1
HTTP	92.8	284.1
FTP	126	286
printer	450	775

This table shows that, for a 55 KB file, the e-mail application uses an average of 157.3 Kbps, the HTTP application uses an average of 92.8 Kbps, the FTP application uses an average of 126 Kbps, and the printer application uses an average of 450 Kbps. The

printer application appears to use much more channel capacity than the other three applications. The 140 KB file size column shows similar results. Although three of the applications appear to be very similar to one another with respect to channel utilization, the printer application could be distinguished from the other three based on utilization alone because of its larger utilization numbers.

Another characteristic of network traffic that could distinguish one application from another is the total number of packets transmitted for a given file size using a given application. Table 4-2 shows the total number of packets used to transmit a given file size for a given application.

Table 4-2 Total Packets Transmitted

	55 KB File	140 KB File
e-mail	373	809
HTTP	279	606
FTP	280	592
printer	1575	3200

For a 55 KB file, the e-mail application uses 373 packets, the HTTP application uses 279 packets, the FTP application uses 280 packets, and the printer application uses 1575 packets. The printer application uses significantly more packets than the other three applications for a given file size. The 140 KB file size column reveals similar results. Knowing the total number of packets used to transmit a file could be used in determining the type of application used in the transmission. Using the total number of packets to determine the application could also be problematic. For example, if a very small file is printed and 500 packets are used to complete that print job, the same number of packets

could also be used to send a moderate sized e-mail. Thus, 500 packets are used for both applications, and determining which application was used becomes very difficult using this technique.

4.2   Final Analysis

The previous sections of this chapter indicate that packet size distributions and histograms can be used to determine what application is the source of a given traffic stream. Additionally, the utilization of the channel can also be an indicator of what application is being used. The total number of packets transmitted is another characteristic studied. This is shown to be problematic when used to distinguish among the applications. The packet size distributions and histograms tend to be a better distinguisher of applications than channel utilization and total number of packets based on the results of the utilization table. This utilization table only distinguishes the printer application from the other three while the packet size distributions and histograms distinguish three and, in some cases, four applications.

The histograms reveal interesting facts about the sizes of the packets used for a given application. Packets of size 0 to 200 bytes tend to be packets consisting of MAC protocol like broadcast, clear to send, ready to send, and acknowledgements. These packets could also be considered noise. An important characteristic about traffic on this ad-hoc network is broadcast packets are sent only when no other application is using the channel. These packets are 61 bytes in size. Therefore, when 61 byte packets are the only packets on the network, the conclusion can be made that the channel is essentially idle. When building the histograms, these broadcast packets make up a portion of the "noise." Packets of size 201 to 900 bytes are also important in distinguishing text based

web pages from image based web pages in HTTP traffic. Packets of size 901 to 1200

bytes are present in all e-mail traffic and can be used to distinguish e-mail from all other

applications. Packets of size 1201 to 1330 bytes are present in HTTP traffic and Internet

browser based FTP traffic. Although a definite conclusion cannot be made that

distinguishes HTTP traffic from Internet browser based FTP traffic, the conclusion can

be made that an Internet browser is open and being used to either access web pages or

transfer files when packets of size 1201 to 1330 bytes are present. Packets of size 1331

to 1450 bytes are present in all printer traffic and can be used to distinguish the printer

application from the other three applications. Packets greater than 1450 bytes represent

the MTU and are present in all applications when the transmitted file size is sufficiently

large. These MTU packets are present in the four applications studied when the file size

is approximately 2 KB or larger. The medium sized packets that distinguish one

application from another are present when the file size reaches approximately 10 KB for

all applications. Table 4-3 shows the percentages of packet sizes transmitted in each

application.

Table 4-3 Packet Size Percentage For Each Application

	0-200	201-900	901-1200	1201-1330	1331-1450	>1450
Email (55K)	0.838	0.003	0.027	0.003	0.003	0.128
Email (140K Text)	0.819	0.001	0.031	0.001	0.001	0.147
Email (140K Image)	0.818	0.001	0.029	0.001	0.001	0.149
HTTP (55K Text)	0.843	0.004	0.004	0.046	0.004	0.100
HTTP (55K Image)	0.841	0.100	0.003	0.014	0.003	0.038
HTTP (140K Text)	0.822	0.003	0.002	0.054	0.002	0.117
HTTP (140K Image)	0.828	0.068	0.001	0.034	0.001	0.068
HTTP (Complex)	0.822	0.060	0.001	0.018	0.001	0.097
FTP (140K)	0.828	0.003	0.002	0.002	0.002	0.164
FTP (646K)	0.816	0.000	0.000	0.000	0.000	0.182
Printer (55K)	0.800	0.075	0.006	0.001	0.056	0.063
Printer (90K)	0.801	0.065	0.005	0.000	0.056	0.073
Printer (140K)	0.803	0.059	0.004	0.000	0.058	0.076
Email - Email	0.774	0.000	0.037	0.002	0.000	0.186
Email - HTTP (Text)	0.816	0.001	0.007	0.046	0.000	0.130
Email - HTTP (Image)	0.804	0.021	0.006	0.046	0.000	0.122
Email - FTP	0.803	0.000	0.006	0.000	0.000	0.191
Email - Printer	0.774	0.063	0.013	0.000	0.049	0.100
HTTP - HTTP	0.818	0.016	0.001	0.054	0.000	0.111

4.3   Summary

This chapter presents and analyzes data from the ad-hoc network. The packet size

characteristic is a good indicator of determining what application produces a given traffic

stream. Channel utilization is another possible indicator; however, it only proves

successful for the printer application. Packet size distributions and histograms are only

useful when the transmitted file size is greater than approximately 10 KB. As the file

size increases for a given application, the histograms and distribution graphs grow more

accurate. The data indicates that the distribution and histogram method of determining

applications is only effective when a transmitted file reaches a minimum size threshold

but remains effective as the file size increases.

V. Conclusions and Recommendations

This chapter describes research contributions, limitations, and recommendations for future research in this area. Section 5.1 presents contributions regarding this research effort, Section 5.2 discusses the limitations of the thesis, and Section 5.3 provides recommendations for future research.

5.1   Research Contributions

The hypothesis of this research is that network traffic characteristics will be found that can be used to identify the application in data exchange over an IEEE 802.11b ad-hoc wireless network without examining the contents of packets or knowing the direction of traffic flow. Distinguishing characteristics of network traffic are visually detected based on the distribution graphs and histograms of packet sizes. The first significant contribution this research makes is demonstrating that there are traffic characteristics that can distinguish one application from another. Building packet size histograms proves to be the most efficient and accurate method for distinguishing one application from another. Based on the packet size distribution and histogram of a given traffic stream, the application accessing the network can be determined. The histogram shown in Figure 5-1 summarizes how applications can be determined based on packet size.

The histogram shows that when packets of size 901 to 1200 bytes are present, the application used is e-mail, when packets of size 1201 to 1330 bytes are present, the application used is either HTTP or Internet browser based FTP, and when packets of size 1331 to 1450 bytes are present, the application used is printer. The packets were grouped as a result of observing many known applications over the network.
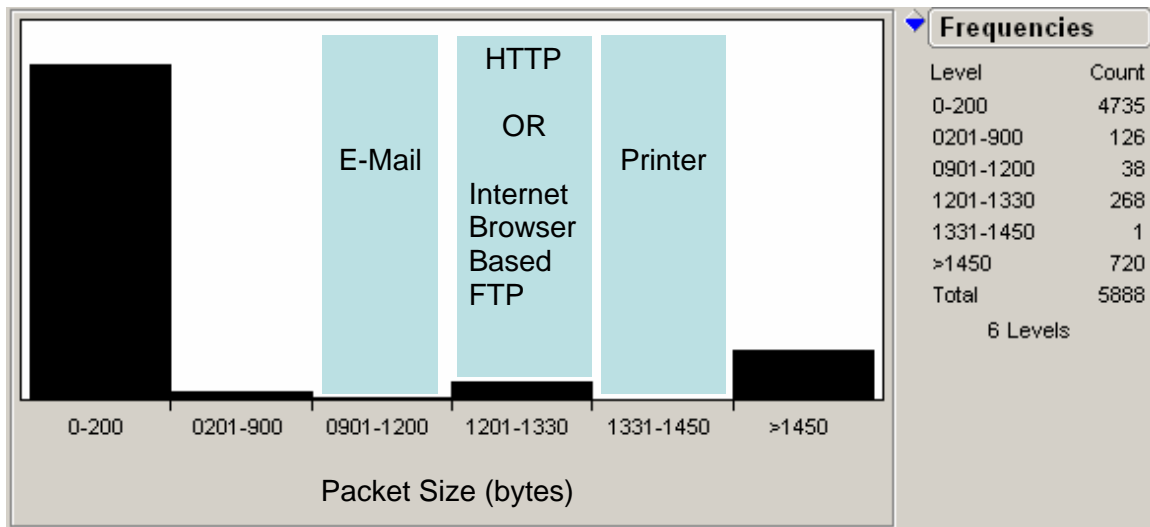
Figure 5-1 Histogram of Packet Sizes of Sample Network Traffic

After observing multiple e-mails, it became clear that all e-mail traffic contains packets sized between 901 and 1200 bytes. This same type of observation is conducted for all applications. In cases where very small files are transmitted via some application over the wireless network, determining the application used becomes difficult with this technique. This is because small files have limited amounts of data, and each application looks similar when graphing and building histograms of the packet sizes. As the transmitted file size increases, the accuracy of determining what type of application is used also increases. Therefore, another significant contribution this research effort provides includes identifying situations where distinguishing among applications over an IEEE 802.11b ad-hoc wireless network becomes problematic.

5.2   Limitations

The data used in this research is collected from an IEEE 802.11b ad-hoc wireless network. This network consists of one server machine, two client machines and one sniffer. Many wireless networks are built with an access point to a wired network,

another wireless network, or the Internet and contain more machines than are present in the research network. For this reason, conclusions drawn from the data collected in the research network do not necessarily generalize to all IEEE 802.11b wireless networks.

A second limitation of this research is the number of applications studied. Although the technique found in this study accurately distinguishes one application from another, this technique may prove less accurate with the addition of other applications onto the network. For example, the packet sizes that characterize printer traffic could also characterize another application not studied in this research effort. In this case, other characteristics or techniques, such as stricter packet size limits, would have to be created in order to potentially distinguish each application.

Another limitation of this research is the dependence on medium sized packets to determine what application is accessing the network. If a network, like ATM, uses fixed length packets, the medium sized characterizing packets are not present, and the histogram technique will be ineffective. If fixed length packets are present, different techniques will need to be determined that characterize the data exchange.

5.3   Recommendations For Future Research

The ad-hoc wireless network used for this research only includes four applications. Many networks have scores of applications being used at any given time. An area of future research includes adding more applications to the network and attempting to characterize data streams using this more complex network. Another area of research would change the version of each application to test the changes in traffic characteristics. For example, a test could be conducted to see if Netscape version 6.0 has different traffic characteristics than Netscape version 7.0.

Future research should add more computers to the network. The data collected in this research results from two computers communicating with one server. As more computers are added to the network, the complexity of determining various applications will undoubtedly increase. Further, each experiment is carefully controlled so that all data from a given application is collected. For example, when e-mail traffic is captured for a given message, all packets from that e-mail transmission are included in the captured data. It is conceivable that in a more complex network, only a small amount of network traffic could be captured. The data will likely be small portions of a few applications and the full transmissions of others. In this case, certain applications could be incorrectly discarded because not enough data would be available for that specific application to be identified in the traffic stream.

Another area of future research includes capturing data from an organizational intranet to test the complexity of the traffic in that environment. As multiple operating systems, applications, server types, microprocessor speeds, network bandwidth speeds, network interface cards, and users are introduced into a network, the characteristics that distinguish one application from another could become more difficult to determine. The goal of the research could be to determine what application is being used by a specific user. If organizational intranet traffic is captured, knowing what user sends or receives that traffic could be very difficult to determine.

Introducing different operating systems in the ad-hoc network is another possibility for future research. The operating system used for this research effort is Windows 2000 and the server is configured using Microsoft Internet Information Services. If a UNIX machine is used as a client or server on the network, the characteristics that distinguish one application from another may change. Another

example for change is using Microsoft Exchange Server for e-mail capabilities rather than the virtual SMTP server used in the current research effort.

Different wireless network cards could be used in the network to test the differences among several wireless network cards. Specifically, the characteristics that distinguish one application from another probably result from the MTU defined by the wireless card used in a given machine. Because all machines in the current research use the same wireless network cards, the MTU is the same for all machines in the network. The vendor specifications state that the MTU can be changed even on the Enterasys cards used in this research [Ent02]. Other types of IEEE 802.11b wireless network cards could have the same capabilities. If the MTU can change size, the medium sized packets that distinguish among applications would probably still be present in the packet size distribution, but the size of those packets would change. For example, the sizes of packets that characterize printer traffic are 1331 to 1450 bytes. If the MTU size increases, more data will fit into a single MTU and the remaining data will possibly be a different size than the size discovered in this research.

In order to study characteristics of network traffic, an ad-hoc wireless network is designed and configured. This appendix outlines the process used to setup, configure, and manage the wireless network used for this research. The setup of the network is accomplished in two phases.

Phase One

The first phase completes the configuration of three laptop computers connected via wireless network cards using a Windows 2000 network workgroup. Because no access point is used in this research, the wireless type is ad-hoc. Another term used for ad-hoc IEEE 802.11b wireless networks is Independent Basic Service Set (IBSS). Both of these terms can be used synonymously, however, this appendix refers to the network as ad-hoc. Figure A-1 shows pictorially how the three computers are setup during phase one of the research.
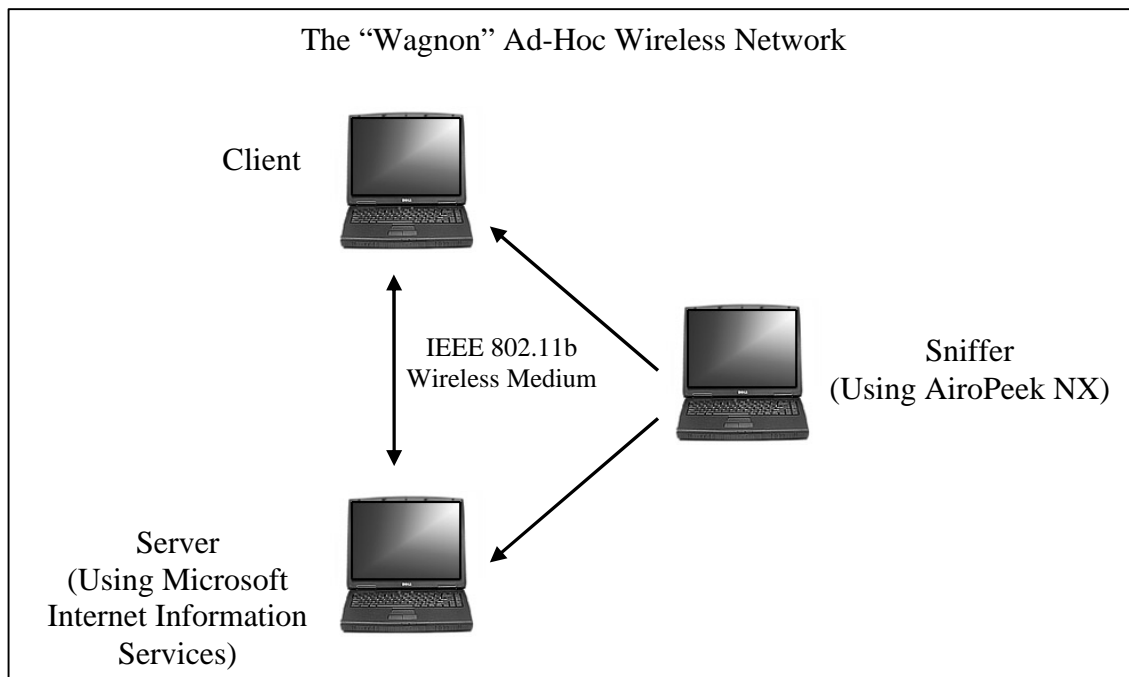

The "Wagnon" Ad-Hoc Wireless Network

Client

IEEE 802.11b
Wireless Medium

Sniffer
(Using AiroPeek NX)

Server
(Using Microsoft
Internet Information
Services)

Figure A-1 Initial Wireless Network Configuration

A-1

The sniffer computer listens in passive mode using a Cisco System air-pcm352 IEEE 802.11b wireless network card with an integrated antenna. This machine is not connected to the network so it simply listens to all machines on the network. The software that captures the wireless packets is AiroPeek NX made by Wildpackets. The operating system on the machine is Microsoft Windows 2000.

The server computer is configured using the Microsoft Windows Internet Information Services (IIS). Microsoft IIS is available as a part of the Windows 2000 operating system. The IIS is accessed under the Services and Applications folder by following the subsequent steps: Start → Settings → Control Panel → Administrative Tools → Computer Management. The IIS allows the administrator to configure the local machine as a server for SMTP, HTTP, and FTP. The printer application is also used but the IIS is not necessary for configuring a remote printer.

E-mail

The virtual e-mail server is configured through the IIS SMTP service. This service allows all e-mail traffic to be sent to the server machine from the client machine. The e-mail vendor used during this research is Microsoft Outlook Express Version 6. The IIS SMTP service only allows for the configuration of SMTP, not POP3. Outlook Express requires the SMTP service to send an e-mail message to the server and the POP3 service to receive e-mail messages from the server. Because the POP3 service is not used, all e-mail traffic is sent from the client machines to the server machine but never from the server machine to the client machine. This is not a problem, however, because the point of the research is to capture the e-mail traffic as it flows from one machine to the other.

HTTP

The HTTP server is configured through the IIS HTTP service. A folder is chosen on the server machine's hard drive to house all web pages and files associated with those web pages such as images and java scripts. This allows all client machines on the network to access the various web pages stored on the server machine. Permissions were set that allowed all clients on the network to access the web pages and images that were included in many of those web pages. The two platforms used by the client machine for accessing the web pages are Microsoft Internet Explorer Version 6.0 and Netscape Version 7.0.

Printer

A virtual Hewlett Packard Laser Jet 8100 series printer is added to the server machine and sharing permissions are set that allowed the client machine access to the printer. There is not an actual printer; however, the applications are able to print to the fictitious printer as if one actually existed. Again, the purpose of the research is to capture the printer traffic as it flows across the wireless medium, so whether the file actually printed on a printer or not is irrelevant.

FTP

The FTP server is configured through the IIS FTP service. This service allows the client machine to send and receive files using various FTP applications. The two platforms used for transferring files are DOS command prompt FTP and Internet Browser FTP using Internet Explorer and Netscape.

The wireless network card used in this machine is the Enterasys Networks csi6d-aa-128 IEEE 802.11b made by Roam About. The operating system used by this machine is Microsoft Windows 2000.

The client machine is used for sending e-mails, printing files, transferring files via FTP, and accessing web pages. The various platforms used for each of these applications are discussed earlier. The wireless network card used in this machine is also the Enterasys Networks csi6d-aa-128 IEEE 802.11b produced by Roam About, and the operating system is Microsoft Windows 2000.

Phase Two

A fourth computer is added to the network in order to accomplish the simultaneous use of two applications. Figure A-2 shows a pictorial representation of the ad-hoc network after the fourth machine is added. The fourth machine is added as a client using the same Enterasys Networks csi6d-aa-128 IEEE 802.11b wireless network card as the other client machine. This second client machine is a desktop, so the wireless network card is inserted into an open PCMCIA slot on the front of the machine. The operating system used on this machine is also Microsoft Windows 2000. This machine, like the other client, is used for sending e-mails, printing files, transferring files via FTP, and accessing web pages. The various platforms used for each of these applications are the same as the other client machine. This client machine is necessary to emulate real network situations by having one client use one application while the other client simultaneously uses another application.
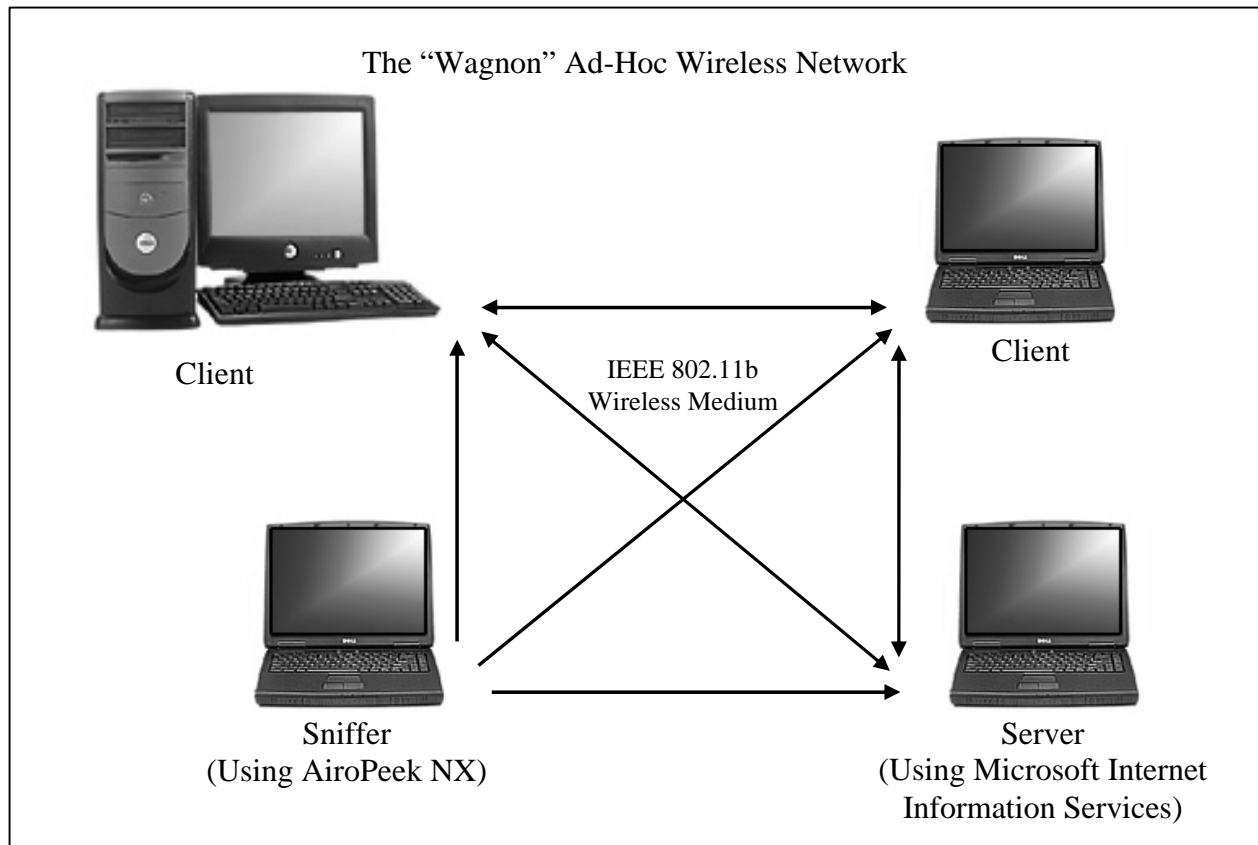
The "Wagnon" Ad-Hoc Wireless Network

Client

Client

IEEE 802.11b
Wireless Medium

Sniffer
(Using AiroPeek NX)

Server
(Using Microsoft Internet
Information Services)

Figure A-2 Secondary Wireless Network Configuration

The workgroup is set up using the Roam About Client Utility and utilizes channel
10 operated at 2.457 GHz [Ent02]. When the Client Utility is loaded onto each of the
computers, the workgroup is set up using the Windows 2000 Network Identification
Wizard. The name of the workgroup is Wagnon. The naming convention is important
because all machines in the workgroup have to use the same workgroup name in order to
successfully join the workgroup [Ent02].

Bibliography

[AgC98]   Agarwal, S. and S. Chaudhuri, "Determination of Aircraft Orientation for a Vision-Based System Using Artificial Neural Networks," *Journal of Mathematical Imaging and Vision*, pp. 255-269, 1998.

[And98]   Andren, C. "IEEE 802.11 Wireless LAN: can we use it for multimedia?" *IEEE Multimedia*, pp. 84-89, 1998.

[BeP92]   Bezdek, J. and S. Pal, *Fuzzy Models For Pattern Recognition: Methods That Search For Structures In Data*, New York, Institute of Electrical and Electronics Engineers, 1992.

[BHM01]   Burns, L., J. Hellerstein, S. Ma, and D. Taylor, "Toward Discovery of Event Correlation Rules," *IEEE/IFIP International Symposium on Integrated Network Management Proceedings*, pp. 345-359, 2001.

[CWK96]   Crow, B., I. Widjaja, J. Kim, and P. Sakai, "Performance of IEEE 802.11 wireless local area networks," *Proceedings of the SPIE - The International Society for Optical Engineering*, pp. 480-91, 1996.

[CWP02]   Claude, I., R. Winzenrieth, P. Pouletaut, and J. Boulanger, "Contour Features for Colposcopic Image Classification by Artificial Neural Networks," *Proceedings 16th International Conference on Pattern Recognition*, pp. 771-774, 2002.

[DaG02]   Dasgupta, D. and F. Gonzalez, "An Immunity-Based Technique to Characterize Intrusions in Computer Networks," *IEEE Transactions on Evolutionary Computation*, pp. 281-291, 2002.

[Ent02]   Enterasys Networks, *802.11 Wireless Networking Guide*, Rochester, NH, 2002.

[GaH00]   Gang, X. and Z. Hui, "Advanced Methods for Detecting Unusual Behaviors on Networks in Real-Time," *IEEE 2000 International Conference on Communication Technology Proceedings*, pp. 291-295, 2000.

[GrR02]   Grabmeier, J. and A. Rudolph, "Techniques of cluster algorithms in data mining," *Data Mining and Knowledge Discovery*, pp. 303-360, 2002.

[HuH97]   Huang, C. and W Huang, "Modified 3-D Hopfield Neural Network for Gesture Recognition," *IEEE International Conference on Neural Networks*, pp. 1650-1655, 1997.

[JDM00]   Jain, A., R. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 4-37, 2000

[Jai91]   Jain, R., *The Art Of Computer Systems Performance Analysis*, New York, John Wiley & Sons, 1991.

[Jam88]   James, M., *Pattern Recognition*, New York, Wiley, 1988.

[JaT99]   Javidi, B. and N. Towghi, "Fully phase encoded techniques for optical security and encryption," *Proceedings of the SPIE - The International Society for Optical Engineering*, pp. 40-56, 1999.

[KaP00]   Kalles, D. and A. Papagelis, "Stable Decision Trees: Using Local Anarchy for Efficient Incremental Learning," *International Journal of Artificial Intelligence Tools*, pp. 79-96, 2000.

[Lak00]   Lakany, H., "A Generic Kinematic Pattern for Human Walking," *Neurocomputing*, pp. 27-54, 2000.

[Mul02]   Mullins, J., "Making Unbreakable Code," *IEEE Spectrum*, pp. 40-45, May 2002.

[PMS00]   Phan, F., E. Micheli-Tzanakou, and S. Sideman, "Speaker Identification Using Neural Networks and Wavelets," *IEEE Engineering in Medicine and Biology Magazine*, pp. 92-101, 2000.

[RuN95]   Russell, S. and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, NJ, 1995.

[SpZ00]   Spafford, E. and D. Zamboni, *Data Collection Mechanisms for Intrusion Detection Systems*, Technical Report 2000-08, Purdue University, 2000.

[Str02]   Stranieri, A., "Discovering Interesting Association Rules from Legal Databases," *Information & Communications Technology Law*, pp. 35-48, 2002.

[TiZ99]   Ting, K. and Z. Zheng, "Improving the Performance of Boosting for Naive Bayesian Classification," *Methodologies for Knowledge Discovery and Data Mining. Third Pacific-Asia Conference*, pp. 296-305, 1999.

[TWF00]   Trafalis, T., A. White, and A. Fras, "Data Mining Techniques for Tornadic Pattern Recognition," *Proceedings of the Artificial Neural Networks in Engineering Conference*, pp. 455-460, 2000.

[WiF00]   Witten, I. and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Francisco, CA, Morgan Kaufmann, 2000.

Vita

John Wagnon was born in Little Rock, AR, in 1976 and graduated from Cabot High School in 1994. He was commissioned as a Second Lieutenant through the Air Force Reserve Officer Training Corps at the University of Arkansas in May 1999, after he completed the degree of Bachelor of Science in Computer Engineering. His first assignment was with the 333rd Training Squadron at Keesler AFB, MS, where he was a student at the Basic Communications Officer Training (BCOT) school. Upon graduation from BCOT, Lt Wagnon was assigned to the 78th Communications Squadron at Robins AFB, GA. He served in all the flights in the squadron as an Aerospace Communications Expert Lieutenant. Some of his jobs included network management, project management, financial advisor, and executive officer. He was selected to attend the Air Force Institute of Technology (AFIT) in November 2000 and reported to Wright Patterson AFB, OH, in August 2001. Lt Wagnon will be assigned to the Air Force Communications Agency at Scott AFB, IL, upon graduation from AFIT.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)* 25-03-2003	2. REPORT TYPE **Master's Thesis**	3. DATES COVERED *(From – To)* Apr 2002 – Mar 2003

4. TITLE AND SUBTITLE CHARACTERIZING DATA STREAMS OVER IEEE 802.11b AD-HOC WIRELESS NETWORKS	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Wagnon, John, T, First Lieutenant, USAF	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way, Building 640 WPAFB OH 45433-7765	8. PERFORMING ORGANIZATION REPORT NUMBER AFIT/GIR/ENG/03-03

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Security Agency Attn: Mr. William Kroah NSA/R5 9800 Savage Road Comm: (301) 688-0348 Ft. George G. Meade, MD 20755-6779 e-mail: wtk@afterlife.ncsc.mil	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Soon, advancements in data encryption technology will make real-time decryption of the contents of network packets virtually impossible. This research anticipates this development and extracts useful information based on packet level characteristics. Distinguishing characteristics from e-mail, HTTP, print, and FTP applications are identified and analyzed. The analysis of collected data from an ad-hoc wireless network reveals that distinguishing characteristics of network traffic do indeed exist. These characteristics include packet size, packet frequency, inter-packet correlation, and channel utilization. Without knowing the contents of packets or the direction of the traffic flow, the applications accessing the wireless network can be determined.

15. SUBJECT TERMS

Computer Networks, Pattern Recognition, Data Processing, Local Area Networks, Electronic Security

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Rusty O. Baldwin, Maj, USAF (ENG)
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	UU	96	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-3636, ext 4612; e-mail: rusty.baldwin@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18